FR-05-26

Human Resources Research Organization

# The Accuracy of School Classifications for the 2004 Accountability Cycle of the Kentucky Commonwealth Accountability Testing System

R. Gene Hoffman
Emily R. Dickinson

Human Resources Research Organization (HumRRO)
950 Breckenridge Lane, Suite 170
Louisville, KY 40207
Phone (502) 721-9045
FAX (502) 721-9983

May 2005

*Prepared for:* Kentucky Department of Education
Capital Plaza Tower, 18th Floor
500 Mero Street
Frankfort, KY 40501

# THE ACCURACY OF SCHOOL CLASSIFICATIONS FOR THE 2004 ACCOUNTABILITY CYCLE OF THE KENTUCKY COMMONWEALTH ACCOUNTABILITY TESTING SYSTEM

## Table of Contents

Introduction ........ 1
    CATS Long-term Accountability Model ........ 1
School Classification Accuracy Results ........ 3
Technical Details for Calculating School Classification Accuracy ........ 8
    Overview ........ 8
        Estimating Standard Errors of Measurement ........ 9
        Note on multiple SEMs ........ 11
        Estimating Classification Accuracy Probabilities ........ 11
    Assessment SEM Computations ........ 12
        Identifying Target School Sizes ........ 13
        Selecting Eligible Schools ........ 14
        Estimating Assessment SEMs using Generalizability Theory ........ 15
        Interpolating Assessment SEMs for School Sizes 10 to 500 ........ 16
    Accountability Index SEMs Computations ........ 17
    Classification SEM Computations ........ 17
        True Target Indexez ........ 18
        True Goal Target ........ 18
        True Assistance Target ........ 18
        Classification and Classification SEM ........ 19
    Shift from Standard Error to Probability Matrices ........ 20
        Creating Probability Matrix of of Observed Differences Given Possible True Differences ........ 20
        Estimating of True Index Variance ........ 20
        Calculating Total Variance in Classification Difference Scores and School Size ........ 21
        Estimating Distributions of True Differences ........ 21
        Creating Matrix of Probabilities of True Differences Given Observed Difference ........ 22
    Summarize Probabilities Across Schools ........ 22
References ........ 24
Appendix Technical Documentations ........ 25
    Generalizability Models ........ 25
        Random, fixed, or sampled from a finite universe ........ 25
        Absolute and relative error ........ 29
        Special Considerations for Writing Assessments ........ 29
        CTBS/5 ........ 30
    Weights used in Calculating accountability index Score and accountability index SEMs ... 37
    Appendix References ........ 38

## List of Tables

Table 1 Classification Probabilities for 2004 School Accountability ........ 4

Table 2 Classification Probabilites for 2004 School Accountability without Baseline Safety Net 5
Table 3 Classification Probabilities with and without SEM Safety Net ........................................ 6
Table 4 Assessments by Grade Level ............................................................................................ 8
Table 5 Achievement Levels and Point Values for Mathematics, Reading, Science, and Social Studies ............................................................................................................................. 9
Table 6 Achievement Levels and Point Values for Arts & Humanities, Practical Living/Vocational Studies, and Writing ............................................................................ 9
Table 7 Identification of Representative School Sizes for Kentucky Core Content Tests ........... 14
Table 8 Identification of Representative School Sizes for CTBS/5 ............................................ 14
Table 9 Ranges of candidate school sizes ..................................................................................... 15
Table A-1 Estimating Variance Components for Pupil: School Year Form Generalizability Theory Design – Random Effects Estimates ..................................................................... 27
Table A-2 Estimating Variance Components for Pupil: School Year Form Generalizability Theory Design – G-Study Estimates................................................................................. 28
Table A-3 Estimating Variance Components for Pupil: School Year Form Generalizability Theory Design – D-study Estimates ................................................................................. 28
Table A-4 Variance Components for Each Grade/Subject By School size Configuration for baseline 1999-2000 ........................................................................................................... 31
Table A-5 Variance Components for Each Grade/Subject By School size Configuration for End-of-Cycle 2003-2004 .......................................................................................................... 34
Table A-6 Weight used in Calculating accountability index Score and accountability index SEMs................................................................................................................................ 37

## List of Figures

Figure 1. Representative School Growth Chart. ..............................................................................2
Figure 2. Schematic representation of matrices used in calculating classification accuracy........12
Figure 3. Classification difference standard deviations for Meeting Goal by school size. ..........21
Figure 4. Classification difference standard deviations for Assistance by school size. ...............21
Figure 5. Classification difference means for meeting goal by school size.................................22
Figure 6. Classification difference means for assistance by school size. ....................................22
Figure A-1. Generalizability theory design representing Kentucky Core Content Test two-year accountability cycle. ..........................................................................................26

# THE ACCURACY OF SCHOOL CLASSIFICATIONS FOR THE 2004 ACCOUNTABILITY CYCLE OF THE KENTUCKY COMMONWEALTH ACCOUNTABILITY TESTING SYSTEM

**R. GENE HOFFMAN**
**LAURESS L. WISE**
**EMILY R. DICKINSON**

**HUMAN RESOURCES RESEARCH ORGANIZATION**

## Introduction

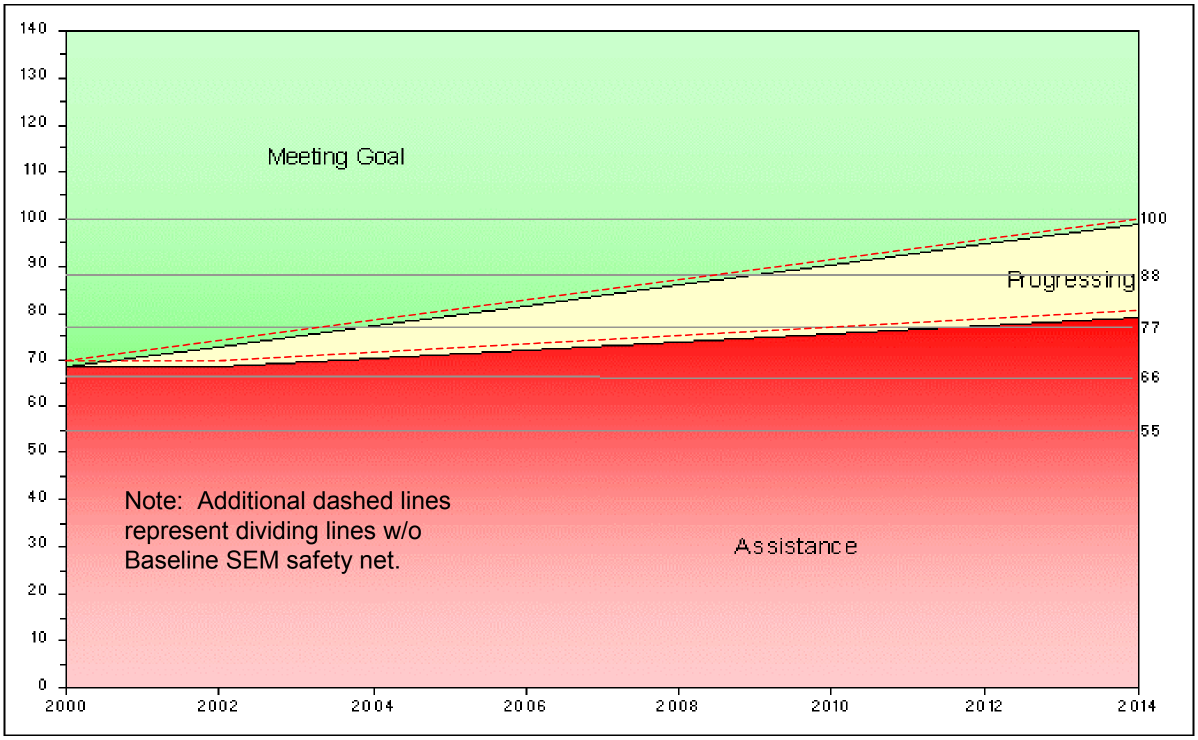
Kentucky's Commonwealth Accountability Testing System (CATS) was implemented in 1999 as a modification of the Kentucky Instructional Results Information System (KIRIS). Beginning with KIRIS, public schools in Kentucky have been classified by their successes in educating students. Both the KIRIS and CATS systems have significant consequences tied to schools' classifications making the accuracy of these classifications an important issue. Hoffman and Wise (2001) reported the accuracy of these classifications for the interim accountability cycle that bridged KIRIS and CATS. The present report presents the method used for calculating classification accuracy and the results for the first of the CATS long-term accountability cycles that are legislated to occur every two years beginning in 2002 and ending in 2014.

The report begins with an overview of the CATS long-term accountability model and then presents classification accuracy results for the first accountability cycle. Details of how the results were obtained then follow. Although the results are reasonably straightforward, computational details are complex and are mainly presented for technical readers.

### CATS Long-term Accountability Model

The CATS long-term accountability cycle began with the school year of 1998-1999, which was the first year in which the newly revised Kentucky Core Content Test (KCCT) was administered. Because CATS testing occurs in the spring of each school year, we reference each year with the spring date only. Data from 1999 and 2000 constituted the "baseline" years upon which target scores for the period through 2014 have been set for every Kentucky school. These targets will be used to place schools into one of three categories: *Meeting Goal*, *Progressing*, and *Assistance*.

For each school, a School Growth Chart (see Figure 1) is constructed to depict school performance targets from 2000 through 2014. A "goal line" is initially plotted from the point on the chart representing a school's academic index for the baseline period and ending at the point that represents an academic index of 100 in the year 2014. The ending point is the statewide goal for all schools in 2014. The line is then adjusted downward to incorporate an allowance for measurement error. That is, the beginning of the line is actually plotted at one




Note: (Edited from a randomly selected school from the KDE website
http://www.kde.state.ky.us/oaa/implement/School_Report_Card/)

Figure 1. Modified School Growth Chart

standard error of measurement (SEM) below the school's calculated index and ends at one SEM below 100. The SEM refers to measurement error in the baseline accountability index.

Every school in Kentucky has a School Growth Chart indicating its prescribed trajectory, but the chart in Figure 1 has been modified from the ones presented to the schools by showing a goal line *without* measurement error allowance. At the end of every two-year accountability cycle, a school's new accountability index is compared to the solid line that divides the Meeting Goal (medium shaded or green if viewed electronically) area from the Progressing area (lighter shaded or yellow if viewed electronically). If the new Index score is at or above the line, then the school is improving close enough to the true-target rate (the dashed line) to be labeled Meeting Goal.

Figure 1 also shows two additional lines on the chart that divide Progressing and Assistance (darker shaded or red) areas. As defined by Kentucky regulation, the Assistance line begins with the baseline academic index at 2000, is sketched horizontally over to the year 2002 and is then extended to the point at the year 2014 representing an accountability index of 80. Like the goal line, the Assistance line used for actual classification is adjusted downward by one SEM. Again, the dashed line in Figure 1 (which is not presented to schools) shows the true line. The solid line that includes the safety net and divides the Progressing and Assistance areas in the chart is used to classify schools.

The distinction between the solid lines plotted on the chart with the built-in safety net and the dashed lines without the safety net is important for later classification accuracy computations. Throughout this report, we will refer to the “safety net” line and the “true” line to maintain this distinction.

Figure 1 is not the complete story for school classification. In addition to the accountability index scores, two additional criteria are applied before a school can be classified as meeting their goal. These criteria include meeting goals for (a) reducing the proportion of Novice students in their schools and (b) staying within maximum limits on the number of dropouts. At this time, neither of these criteria is considered in this analysis of school classification accuracy.

## School Classification Accuracy Results

No assessment system is perfect, which means that an observed score, such as a school accountability index, is the product of two factors: true standing and measurement error. Although observed scores are known, true scores are not because the exact error in any given score is uncertain. Test reliability statistics, however, allow the estimation of how errors are distributed, making it possible to address the following two questions:

- What is the probability that a school is classified accurately? That is, what is the probability that a schools true scores places the school in the same accountability classification as the one assigned by its observed index scores?
- What is the probability that a school is incorrectly classified? That is, what are the odds that a school’s true scores would result in the school being placed in a different accountability classification from the one assigned?

Table 1 presents a summary of classification accuracy results. The columns indicate school classifications considering only their accountability index scores. Ignored are special criteria concerning reduction in percentages of students classified as Novice and limits on school dropouts. *Italicized numbers* represent percentages of all students, so that their sum is 100% (within rounding). The ***bold italicized numbers*** represent the percent of schools expected to have true scores in a range that would yield the same accountability classification as the assigned classification. Thus, 45% of all schools were assigned “Meeting Goal” and are expected to have true classifications of “Meeting Goal.” Another 34% of all schools were assigned “Progressing” and are expected to have true classifications of “Progressing.” Finally, 3% of all schools are assigned Assistance and are expected to have true classifications of “Assistance.” The sum of the bold percentages, 82%, is the percentage of all schools whose true classifications are expected to match their assigned classifications. That is, school classification accuracy, for the system as implemented, is 82%.

Table 1
Classification Probabilities for 2004 School Accountability

| Expected True Category | Assigned Category (Before Novice and Drop Criteria Applied) | | | Total Expected for True Classifications |
|---|---|---|---|---|
| | Meeting Goal | Progressing | Assistance | |
| Meeting Goal | ***45%*** | *1%* | *0%* | 46% |
| Progressing | *11%* | ***34%*** | *1%* | 46% |
| Assistance | *2%* | *3%* | ***3%*** | 8% |
| % in Observed Class | 58% | 38% | 4% | 100% |
| Number in Obs Class | 694 | 461 | 47 | 1202 |

Notes: Bold italics numbers indicate expected probabilities of accurate classifications. They sum to 82%. Only schools with data for all four years and with constant grade configurations are include in the analysis.

The bottom two rows of Table 1 show the percent of schools and total number with accountability index scores in each observed classification. In the right-most column, the table shows the percent of schools that would be expected in each classification if their true scores were knowable. Notice that more schools are actually assigned to the Meeting Goal category than are expected from our projections about true scores (58% vs. 46%). Conversely, fewer schools are assigned Assistance than are expected (4% vs. 8%). Part of this difference is the result of the application of the baseline safety net: Schools just under their true Goal or Assistance line are given the “benefit of the doubt” via the SEM allowance. As a result, the system places more schools into the Meeting Goal category than expected, but limits the chances that schools are classified too low because of measurement error.

Table 2 shows how accurately the accountability system would be if schools were classified without the baseline SEM safety net. These results are perhaps a better indication of measurement accuracy. Without the safety net, schools would be assigned to the category most likely to contain their true score. Therefore, overall accuracy, at 89% (the sum of the bold percentages in Table 2), is higher without the SEM safety net than with it. While seemingly paradoxical, this result was expected. Including the baseline safety net increases the total number of schools that are classified as Meeting Goal in order to reduce the risk of erroneously under-classifying schools. The result is that some schools are over-classified.

Table 2
Classification Probabilites for 2004 School Accountability without Baseline Safety Net

| Expected True Category | Observed Category Without Applying Baseline SEM Safety Net: Meeting Goal | Progressing | Assistance | Total Expected for True Classifications |
|---|---|---|---|---|
| Meeting goal | ***44%*** | *2%* | *0%* | 46% |
| Progressing | *5%* | ***38%*** | *3%* | 46% |
| Assistance | *0%* | *1%* | ***7%*** | 8% |
| % in Observed Class | 49% | 41% | 10% | 100% |
| Number in Obs Class | 586 | 499 | 117 | 1202 |

Notes: Bold italics numbers indicate expected probabilities of accurate classifications. They sum to 89%.
Only schools with data for all four years and with constant grade configurations are include in the analysis.

Table 3 gives a more comprehensive picture by specifically identifying schools that benefited from the baseline safety net. In this table, six types of schools are identified:

1. Schools that are Meeting Goal with and without the baseline safety net (i.e., above the dashed goal line in Figure 1). These are labeled “MG & MG” in Table 3
2. Schools that are Meeting Goal with the safety net, but are Progressing without the safety net (i.e., schools between the solid and dashed goal lines). These are labeled “MG & P” in Table 3.
3. Schools that are Progressing with and without the baseline safety net for the Assistance line. These schools are below the solid goal line and above the dashed Assistance line in Figure 1 and are labeled “P & P” in Table 3.
4. Schools that are in Assistance with and without the baseline safety net (i.e., below the solid assistance line in Figure 1). These schools are labeled “A & A” (last column) in Table 3.
5. Schools that are Progressing with the safety net, but Assistance without the safety net (i.e., between the solid and dashed assistance lines). There schools are labeled “P & A.”
6. Schools that are in Assistance without the safety net, but are Meeting Goal with the safety net. These schools are labeled “MG & A.”

Table 3
Classification Probabilities with and without SEM Safety Net

| Expected True Category | Classification with SEM Safety Net & without SEM Safety | | | | | |
|---|---|---|---|---|---|---|
| | MG & MG | MG & P | P & P | MG & A | P & A | A & A |
| Meeting Goal | **89%** | **21%** | 1% | **0%** | 0% | 0% |
| Progressing | 11% | 79% | **95%** | 14% | **45%** | 14% |
| Assistance | 0% | 0% | 4% | 86% | 55% | **86%** |
| % in Obs Class | 100% | 100% | 100% | 100% | 100% | 100% |
| Number in Obs Class | 586 | 86 | 413 | 22 | 48 | 47 |

MG & MG = Meeting Goal with or without safety net.
MG & P = Meeting Goal with safety net but Progressing without it.
P & P = Progressing with or without safety net.
MG & A = Meeting Goal with safety net but Assistance without it.
P & A = Progressing with safety net but Assistance without it.
A & A = Assistance with or without safety net.

Table 3 shows percentages that total 100 within each column. The values express the likelihood of a given type of school having true index values that would result in a classification of Meeting Goal, Progressing, or Assistance. For example, schools above the dashed goal line (the "MG & MG" schools) have an 89% probability of being accurately classified as Meeting Goals and only a 11% probability of being truly in the Progressing category.

Comparing the "MG & P" to the "P & P" schools shows the effect of applying the safety net more explicitly. Again, the "MG & P" schools are those with index scores categorizing them as Progressing were it not for the safety net. These schools are most likely to have true scores that would place them in Progressing range (79%), but to protect the 21% that are likely to be in the true Meeting Goals range, all of these schools are classified as meeting their goals. That is, in order to avoid under-classifying 22% of these schools, 79% (100%-21%) of them are over-classified. In contrast, those schools below the solid goal line and above the solid assistance line (the "P & P" schools that are progressing with or without the safety net) have a 95% chance of truly being Progressing and only a 1% chance of truly being Meeting Goal. In other words, if a school has received a classification of Progressing, the odds are high that the school's true standing, if known, would be in Progressing.

Schools in the final three columns all have scores that place them in Assistance without the safety net and in each case the probabilities are greater for them being in Assistance than any other category. In all three cases, there is no chance that a school classified as in Assistance without the safety net would actually be Meeting Goal. The "P & A" schools have a 45% chance of actually being Progressing and are granted this status by the safety net. "A & A" schools (assistance by both classifications) have an 86% probability of being accurately classified. For the schools that were classified as needing Assistance, chances are high that the classification is accurate.

Note that the safety net had to be set prior to the availability of complete data for 1999 through 2002 and was chosen to be one SEM in the baseline accountability index. The actual error is a function of measurement error in both baseline and end-of-cycle scores. The data in

Table 3, therefore, indicate how well the safety net actually protected schools from being misclassified. It seems to have functioned quite well in protecting individual schools from under classifications by measurement error.

These adjusted assignments may not be the best way to view the state learning progress as a whole. The safety net assignments indicate that 58% of schools are meeting their accountability goals. On the other hand, the expected true distributions (last column in Table 1 or 2) indicate that, if measurement error were removed, only about 46% of the schools meet the intended growth targets. A better estimate of state-wide school improvement is provided by the proportion of schools which would have been classified as Meeting Goal without the safety net factor (49%, according to Table 2). Note that this is up from a similar estimate of 40% (Hoffman & Wise, 2003) at the end of the 2002 accountability cycle.

A comparison between 2004 and 2002 shows some significant improvements in the classification process. As illustrated in Table 1 (e.g. School classifications prior to Novice and Drop criteria being applied), 45% of schools were classified as Meeting the Goal when they were expected to meet the goal, a significant increase from 2002 (34%). Overall, significantly more schools are classified as Meeting Goal in 2004 (58%) than in 2002 (49%). Without applying the Safety Net, significantly more schools are classified as Meeting Goal when expected to Meet Goal in 2004 (44%) than in 2002 (32%). Overall, more schools are being classified as Meeting Goal in 2004 (49%) than in 2002 (40%) while those being classified as Progressing has decreased in 2004 (41%) from 2002 (47%). Finally, when comparing classifications with and without the Safety Net, significantly more schools are being classified as meeting goal with and without the Safety Net in 2004 (89%) than in 2002 (80%). Also, significantly more schools are being classified as Progressing with and without the Safety Net in 2004 (95%) than in 2002 (90). Table 4 shows that overall accuracy has increased from 2002.

Table 4. Total Percent of Schools Correctly Classified

| | Year | Percent * |
|---|---|---|
| Without SEM Safety Net | 2002 | 82% |
| | 2004 | 89% |
| With SEM Safety Net | 2002 | 77% |
| | 2004 | 82% |

*Results indicate that the percentage of schools correctly classified has significantly increased between 2002 and 2004 when looking at both classification percentages both with and without the safety net.

Technical Details for Calculating School Classification Accuracy

The material that follows is technical in nature because of the large number of steps involved in reaching the results. This section is written for the technical audience. Some of the steps are straightforward. Other steps require the technical reader to think in some unusual ways. Much of this complexity is created by the need to consider the set of Kentucky schools not as a single population (which is normally the case when considering test statistics), but as representing multiple populations with measurement characteristics that differ by school size and by school grade configuration. An additional complication is that schools were classified based on index differences from both goal line projections and assistance line projections. The presentation begins with an overview of the procedure and then unfolds with details of the computations.

## Overview

Student-level Kentucky Core Content Test scores are used to compute school accountability index scores. These tests are administered to selected grades such that all assessments are administered in typical elementary, middle, and high schools. Eight assessments are components of the KCCT and are prepared for Kentucky to assess achievement.[1] The eight assessments are augmented by a national norm-referenced test, the CTBS/5. Table 4 indicates the grades in which the assessments are administered. Kentucky Core Content Tests are indicated by subject.

Table 4
Assessments by Grade Level

| Subject | Grade | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| Arts & Humanities | | | X | | | X | | | X | |
| Mathematics | | | X | | | X | | | X | |
| Practical Living/Vocational Studies | | | X | | | X | | X | | |
| Reading | | X | | | X | | | X | | |
| Science | | X | | | X | | | | X | |
| Social Studies | | | X | | | X | | | X | |
| On-demand Writing Prompt | | X | | | X | | | | | X |
| Writing Portfolios | | X | | | X | | | | | X |
| CTBS/5 | X | | | X | | | X | | | |

For each KCCT, students are classified into one of four achievement levels: Novice, Apprentice, Proficient, and Distinguished. The lower two levels, Novice and Apprentice, are subdivided into three sublevels (low, middle, and high) for of the four primary content disciplines (Reading, Mathematics, Science, and Social Studies). The point values used to calculate schools' average student achievement for primary content areas are shown in Table 5 and other areas in Table 6.

[1] As defined by the Kentucky Core Content Assessment and laid out by the Kentucky Core Content Test Blueprint (**http://www.kde.state.ky.us/oaa/valid/blueprint.asp**).

Table 5
Achievement Levels and Point Values for Mathematics, Reading, Science, and Social Studies

| Achievement Level | | Point Value |
|---|---|---|
| Distinguished | | 140 |
| Proficient | | 100 |
| Apprentice | High | 80 |
| | Middle | 60 |
| | Low | 40 |
| Novice | High | 26 |
| | Middle | 13 |
| | Low | 0 |

Table 6
Achievement Levels and Point Values for Arts & Humanities, Practical Living/Vocational Studies, and Writing

| Achievement Level | | Point Value |
|---|---|---|
| Distinguished | | 140 |
| Proficient | | 100 |
| Apprentice | | 60 |
| Novice | Attempt | 13 |
| | No attempt | 0 |

CTBS/5 scores are included in the school accountability formula by converting percentiles to a scale similar to that for the KCCT. Specifically, student's quartiles (lowest to highest) are converted to scores of 0, 60, 100, and 140. These scores are used to compute schools' average CTBS/5 scores.

In addition to the KCCT and CTBS/5 data, schools also receive scores for a composite of non-academic factors such as attendance rate, retention rate, and dropout rate. Each school generates the non-academic data.

Given this array of data, estimating school classification accuracy can be conceptualized as a two-phase process that begins with the estimation of SEMs, or error variance, for schools' accountability cycle scores and is followed by transformation of error variance into the classification accuracy probabilities that appear in Tables 1, 2, and 3.

*Estimating Standard Errors of Measurement*

Schools' achievements are classified for CATS based on the difference between their end-of-cycle targets and their end-of-cycle accountability indexes. Therefore, the measurement error of most interest is the error in this difference. Error in the difference, however, is a function of the error in the baseline index (which is used to compute end-of-cycle targets) and the error in the end-of-cycle index. The estimation of these errors is complicated by a variety of factors.

First, school accountability index scores, for any cycle, are a weighted composite (weighted sum) of the scores from the various assessments administered in the schools. Therefore, the SEM for each accountability index (baseline and end-of-cycle) can be computed from SEMs for each assessment used in the computation (i.e., the KCCTs, CTBS/5, and non-academic indicators). As a result, the analyses deal with three types of SEMs:

- **Assessment SEMs** for school-level scores for Grade 4 Reading, Grade 10 Reading, Grade 9 CTBS/5, etc.

- **Accountability Index SEMs** for the baseline school index and each end-of-cycle index. Accountability SEMs are a function of assessment SEMs.
- **Classification SEMs,** which indicate the measurement error in the difference between observed accountability index and the goal for any particular accountability cycle. Classification SEMs are a function of accountability SEMs.

Generalizability Theory analyses, modeled on Yen (1997) and Miller (1999), are used to calculate assessment SEMs for the all except the non-academic indicators. The Generalizability analyses are identical to those used in calculating classification accuracy for the interim accountability model. Two Generalizability models were used: one for KCCTs with different forms in a given year and one for assessments in which all students had the same form. Details of these analyses are presented in Hoffman and Wise (2000b and 2001) and are repeated in the Technical Appendix of this report. In general, the model considers student scores as data points (in lieu of test items) but it is complicated by the fact that school scores for any end-of-cycle assessment are derived from different students for the two years in the cycle with these students taking multiple forms of the assessments that also differ across years. In other words, each student is like a test item, providing a single measure of the instructional capacity of the school. The test item analogy, however, is complicated by the two-year measurement period and by the potential for differences in test forms to impact how students function as a yardstick of school capacity. Variations in student, forms, and years can signal potential sources of measurement error. Further discussion is provided in the Appendix.

No method existed for estimating the error variance for the non-academic scores, so when computing classification accuracies for the interim accountability model Hoffman and Wise (2001) explored using the SEM values based on an assumed reliability of 1 (perfect reliability) and values based on an assumed reliability of 0 (total unreliability). It was determined that the estimate of overall school error was only slightly different for these two extreme assumptions. Therefore, we selected a conservative reliability estimate (.7) for the non-academic scores to use in calculations of school classification accuracy.

The second factor considered in estimating measurement error is the amount of data available for a particular school. Other things being equal, with more data there is less error. As a result of this principle, we expected large schools to be measured more accurately than small schools because their index scores are based on more students. Therefore, analyses of assessment SEMs were conducted on three representative school sizes: the lower third, the middle third, and the upper third.

These considerations mean that for any given cycle there are 81 assessment SEMs estimated by the Generalizability analyses: the 27 grade by assessment content areas (listed in Table 4) times the 3 school sizes.

A third consideration when estimating SEM is the fact that not all schools fit the typical elementary, middle, and high school model. In fact, accountability index SEMs had to be calculated for schools with 14 different grade configurations. (The exact combinations are presented later in Table A-6 in the Appendix.) Fortunately, accountability index SEMs are computed from the separate grade/subject assessment SEM. Therefore, calculating accountability SEMs for schools with any particular grade configuration means including assessment SEMs for the assessments administered in the grades included in that configuration.

A fourth consideration is the requirement to estimate SEMs for a broad range of school sizes. In order to increase the precision of assessment SEM estimates for schools that do not fall in the representative sizes, an interpolation procedure was required to generate assessment SEM estimates for schools with anywhere from 10 to 500 students per grade.

Finally, schools were classified according to how their end-of-cycle accountability index fell in relation to their goal and assistance line targets for that cycle. Therefore, measurement error in the baseline and the end-of-cycle indexes were jointly considered. Computing classification accuracy involves consideration of the differences between a school's actual end-of cycle index and the values specified by that school's true goal and assistance lines, i.e., when the lines are unadjusted by the baseline safety net. Carefully notice that schools actually will be classified according to where their accountability index falls in relation to the goal and assistance lines as plotted to include allowance for measurement error. For purposes of determining classification accuracy, however, schools' end-of-cycle accountability index must be compared to goal and assistance lines that are not adjusted for the potential error. In a sense, the classification accuracy analysis determines the extent to which the error allowance is protecting schools from inappropriately low classifications.

*Note on multiple SEMs*

Because of the complexity of the analysis process with its multiple levels of SEMs (assessment, accountability, and classification), it is easy to lose sight of the fact that within each of these levels, multiple SEMs are computed for varying school sizes and grade configurations. This is much more complex than calculating SEMs for a typical "test" in which one given set of observations (e.g., test items) in the assessment is the same for all subjects. In the case of school assessment, the number of observations in the assessment procedure depends on the number of grades in a school and the number of students within those grades. As a result, every school size and grade configuration combination has a specific set of assessment SEMs. Likewise, every school size and grade configuration combination has a specific accountability SEM. Finally, classification SEMs depend on school size in the baseline year, school size at the end-of-the cycle, and school configuration. Our classification SEM computations allow school size to change. On the other hand, a change in grade configuration invokes special regulations, typically involving the use of district-level scores. Therefore, our classification SEM computations exclude schools that change grade configurations.

*Estimating Classification Accuracy Probabilities*

Standard errors of measurement indicate expected variations of observed scores given a particular true score. Schools, however, have only their observed scores and are interested in how their true score might vary from their observed score. Our method for calculating classification accuracy is based on obtaining estimates of the distribution of true scores around observed scores. In our analyses of student classification accuracy (Hoffman and Wise, 2000a) and interim accountability classification accuracy (Hoffman and Wise, 2001), we applied

Bayes' Theorem and estimates of true score distributions to transform SEMs into estimates of the distribution of true scores around varying levels of observed scores.

Figure 2 illustrates the steps. First, classification SEMs are used to construct a matrix of the probabilities of observed scores given various possible true differences. The figure illustrates that these calculations are made for score intervals spreading from 0 in increments of .5 for possible true scores and observed scores. Using estimates of the probabilities of the various true scores, the top matrix in Figure 2 is converted to the bottom matrix of probabilities of various true score given potential observed scores. These operations are conducted twice: Once for differences around the goal line and once for differences around the assistance line. The shaded area in the lower table identifies accurate classifications.

| Possible True Difference | Possible Observed Difference | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | <= etc. | -1.5 to -1.0 | -1.0 to -.5 | -.5 to 0 | 0 to .5 | .5 to 1.0 | 1.0 to .5 | etc. => |
| <= etc. | | | | | | | | |
| -1.5 to -1.0 | | | | | | | | |
| -1.0 to -.5 | | | | | | | | |
| -.5 to 0 | | | | | | | | |
| 0 to .5 | | | | | | | | |
| .5 to 1.0 | | | | | | | | |
| 1.0 to .5 | | | | | | | | |
| etc. => | | | | | | | | |

Prob(Obs|True)

Via Prob(True)

| Possible True Difference | Possible Observed Difference | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | <= etc. | -1.5 to -1.0 | -1.0 to -.5 | -.5 to 0 | 0 to .5 | .5 to 1.0 | 1.0 to .5 | etc. => |
| <= etc. | | | | | | | | |
| -1.5 to -1.0 | | | | | | | | |
| -1.0 to -.5 | | | | | | | | |
| -.5 to 0 | | | | | | | | |
| 0 to .5 | | | | | | | | |
| .5 to 1.0 | | | | | | | | |
| 1.0 to .5 | | | | | | | | |
| etc. => | | | | | | | | |

Prob(True|Obs)

Figure 2. Schematic representation of matrices used in calculating classification accuracy.

Using the matrix of probabilities of true scores given observed scores, we can sum cells above and below 0 (the differently shaded areas) to estimate classification accuracy. The first step is to identify the column that contains a school's observed difference. Using the goal difference matrix, the probability of the school having a true meeting goal classification is the sum of the values in the identified column that are above the 0 true score. Using the assistance difference matrix, the probability of the school having a true assistance classification is the sum of the values in the identified column that are below the 0 true score. Since each school's true classification must be in Meeting Goal (MG), Assistance (A), or Progressing (P), the probability of the school's true classification being progressing can be calculated as:

Prob(P|Observed Index) = 1-Prob(MG|Observed Index) -Prob(A|Observed Index).

## Assessment SEM Computations

Assessment SEM are derived from Generalizability Theory analyses modified by a four-step process to consider varying school sizes:

- Identify representative target school sizes for Generalizability Theory analysis
- Create synthetic schools with target sizes
- Compute Generalizability Theory error estimates
- Interpolate assessment SEMs for school sizes 10 to 500 per grade.

Each is discussed in detail below.

*Identifying Target School Sizes*

The number of students within a school will affect the reliability of school-level scores; therefore, we begin assessment SEM computations using three representative school sizes. Because schools also differ in the number of grades they contain, and because the analysis begins with grade-level data, we defined school size by the average number of students in a grade. Small schools were identified as those in the smallest one-third of all schools, and the representative size was set at the median of that third, which is the 16.7$^{th}$ percentile of all schools. Similarly, medium size schools were those in the middle one-third and were represented by the 50$^{th}$ percentile of all schools. Finally, large schools were the largest one-third and were represented by the 83.3$^{rd}$ percentile of all schools.

The selection of representative school sizes was slightly complicated by the requirement to analyze data from different grades for two different years. That is, either the grade-level size for 2003 or 2004, or an average, could define school percentiles. Representative size was also affected by test form configuration. The KCCT is divided into multiple forms and we needed each form to be represented by an equal number of students in our analyses. Therefore, target sizes had to be adjusted to the nearest multiple of 12, which is the number of Arts & Humanities and Practical Living/Vocational Studies forms. By using 12 as the multiple, we also accommodated the 6 forms for the other subject areas.

Table 7 below shows the distribution of school size by grade and year, as computed for the KCCT during analyses of interim accountability classification accuracy. School sizes during the interim accountability cycle were the same as during the initial two years of the long-term accountability cycle. Therefore, school size targets determined for the interim classification accuracy analyses are usable for the present analysis. For reference, school sizes at the medians and the boundaries of the one-third size divisions are indicated, along with the maximum school size. Although there are 14 grade configurations for which accountability SEMs are calculated, schools with Grade 4 always include Grade 5, schools with Grade 7 always include Grade 8, and Grades 10, 11, and 12 are always combined. Hence, school size targets were set for Grade 4 and 5, Grade 7 and 8, and High School for the Kentucky Core Content Test. High School targets were set using only population data for Grade 10 and 11. We used these same school size targets when calculating end-of-cycle assessment SEMs because (1) school populations were not expected to shift sufficiently within the need to target a multiple of 12, and (2) an interpolation procedure was applied to cover the range of school sizes

Table 7
Identification of Representative School Sizes for Kentucky Core Content Tests

| Grade | Year | School Sizes by Percentile | | | | | |
|---|---|---|---|---|---|---|---|
| | | 16.7th | 33.3rd | 50th | 66.7th | 88.3rd | Maximum |
| 4 | 2003 | 30 | 45 | 59 | 75 | 96 | 246 |
| 4 | 2004 | 29 | 47 | 61 | 76 | 96 | 255 |
| 5 | 2003 | 28 | 44 | 57 | 73 | 89 | 290 |
| 5 | 2004 | 30 | 46 | 59 | 75 | 94 | 291 |
| Grade 4/5 targets | | **24** | | **60** | | **96** | |
| 7 | 2003 | 35 | 70 | 126 | 191 | 246 | 438 |
| 7 | 2004 | 36 | 67 | 127 | 190 | 259 | 459 |
| 8 | 2003 | 36 | 71 | 133 | 191 | 256 | 430 |
| 8 | 2004 | 36 | 70 | 126 | 194 | 247 | 423 |
| Grade 7/8 targets | | **36** | | **120** | | **240** | |
| 10 | 2003 | 61 | 115 | 179 | 228 | 298 | 624 |
| 10 | 2004 | 63 | 119 | 173 | 222 | 292 | 644 |
| 11 | 2003 | 65 | 110 | 164 | 202 | 258 | 563 |
| 11 | 2004 | 65 | 110 | 163 | 206 | 261 | 518 |
| High School target | | **60** | | **168** | | **240** | |

Table 8 presents targets for CTBS/5 grades derived the same way as described above.

Table 8
Identification of Representative School Sizes for CTBS/5

| Grade | Year | School Sizes by Percentile | | | | | |
|---|---|---|---|---|---|---|---|
| | | $16.7^{th}$ | 33.3rd | 50th | 66.7th | 88.3rd | Maximum |
| 3 | 2003 | 31 | 47 | 63 | 79 | 105 | 275 |
| 3 | 2004 | 31 | 46 | 62 | 80 | 106 | 254 |
| Grade 3 targets | | **24** | | **60** | | **96** | |
| 6 | 2003 | 25 | 40 | 59 | 98 | 222 | 449 |
| 6 | 2004 | 25 | 40 | 59 | 101 | 228 | 383 |
| Grade 6 targets | | **24** | | **60** | | **180** | |
| 9 | 2003 | 33 | 103 | 173 | 244 | 356 | 643 |
| 9 | 2004 | 61 | 113 | 194 | 249 | 365 | 590 |
| Grade 9 targets | | **60** | | **168** | | **240** | |

*Selecting Eligible Schools*

Given that there are not schools with exactly the target number of students nor with an equal representation of forms, we created synthetic schools to match the targets. This was done by randomly eliminating students from candidate schools. Because small, medium, and large size schools have characteristics other than size that may affect measurement accuracy (e.g., smaller schools may be more homogeneous), only schools near the target size were considered eligible for the analyses. Certainly, schools could be no smaller than the target size. Selection of the maximum size eligible for the analysis was a trial and error process. In each case, we tried to balance having enough schools for stable Generalizability results without having the maximum size being subjectively larger than the target size. This was most difficult to achieve for the small middle and high schools. Random selection of students was conducted independently for every grade, subject, and school size combination. Table 9 indicates the ranges of school sizes (target to maximum) that became candidates. The numbers of schools that met each criterion and were used in the Generalizability Theory analysis are presented later.

Table 9
Ranges of candidate school sizes

| | Small | | Medium | | Large | |
|---|---|---|---|---|---|---|
| Grade | Target Size | Max. Size | Target Size | Max. Size | Target Size | Max. Size |
| 3 | 24 | 36 | 60 | 72 | 96 | 120 |
| 4 | 24 | 36 | 60 | 72 | 96 | 120 |
| 5 | 24 | 36 | 60 | 72 | 96 | 120 |
| 6 | 24 | 36 | 60 | 84 | 180 | 240 |
| 7 | 36 | 60 | 120 | 172 | 240 | 360 |
| 8 | 36 | 60 | 120 | 172 | 240 | 360 |
| 9 | 60 | 120 | 168 | 240 | 240 | 643 |
| 10 | 60 | 120 | 168 | 240 | 240 | 643 |
| 11 | 60 | 120 | 168 | 240 | 240 | 643 |
| 12 | 60 | 120 | 168 | 240 | 240 | 643 |

*Estimating Assessment SEMs using Generalizability Theory*

After creating synthesized schools at the target student populations, assessment SEMs were calculated using the Generalizability models specified by Hoffman and Wise (2000b, 2001) and repeated in the Appendix. Results for baseline years appear in Table A-4 and the 2002 end-of-cycle results are in Table A-5. The assessment SEMs required for computation of accountability index SEMs are the square roots of the Generalizability Theory absolute error variance estimates. Absolute error was chosen because schools must meet fixed standards. Relative error is inappropriate because making comparisons to other schools does not play a role in classifying schools. Tables A-4 and A-5 also provide other Generalizability results, including relative error variance, total variance, and absolute and relative Generalizability coefficients. The Generalizability coefficients estimate the reliabilities of the school mean test scores for each assessment included in CATS. In general, these reliabilities are in the mid-eighties to mid-nineties and are higher for the larger schools than the smaller schools.

To estimate error variance for the non-academic component of the accountability index, total variance across schools (separately for elementary, middle, and high schools) was calculated and multiplied by 1 minus our assumed reliability of .7. The square root of that result yielded our estimate of non-academic SEM. The same non-academic SEM is used for all school sizes, because normal measurement theory may not apply. That is, large school may have a more difficult time getting accurate data about each of their students than small schools that may counteract the general measurement principle that more data decreases measurement error.

*Interpolating Assessment SEMs for School Sizes 10 to 500*

In the previous step, assessment SEMs were produced for representative school sizes. In order to increase the precision of the SEMs for schools with student populations at other than the representative sizes, an interpolation procedure was used for each grade/assessment combination. This procedure estimated SEMs for school sizes between 10 and 500 by weighting the distance between any given school size and the representative sizes. More specifically, for each assessment the procedure began with the Generalizability absolute error estimates for the three representative school sizes (small, medium, and large), then:

- For each grade-level (g), assessment (a), and representative size (r), within-school, student-level, error standard deviation (SESD) was estimated from the school-level Generalizability Theory absolute error (AERR), number of forms for the assessment (NF), and number of persons per form (NP) for the representative school size (where NF times NP is representative school size = NRS) and the formula relating variance of means (school scores in this case) to variance of observations (students in this case):

$$SESD_{gar} = \sqrt{AERR_{gar} \times NRS_{gar}} \tag{1}$$

- Interpolate within-school error standard deviations for alternate school sizes (or $SESD_{gan}$, where $n$ stands for an alternate size), where $s$, $m$, and $l$ refer to small, medium, and large representative sizes, respectively:

$$\text{If } n \le NRS_s \text{, let } SESD_{gan} = SESD_{gas} \tag{2}$$

$$\text{If } NRS_s < n < NRS_m \text{, let} \tag{3}$$
$$SESD_{gan} = \left(\left(n - NRS_s\right) \times SESD_{gam} + \left(NRS_m - n\right) \times SESD_{gas}\right) \div \left(NRS_m - NRS_s\right)$$

$$\text{If } NRS_m \le n < NRS_l \text{, let} \tag{4}$$
$$SESD_{gan} = \left(\left(n - NRS_m\right) \times SESD_{gal} + \left(NRS_l - n\right) \times SESD_{gam}\right) \div \left(NRS_l - NRS_m\right)$$

$$\text{If } n \ge NRS_l \text{, let } SESD_{gan} = SESD_{gal} \tag{5}$$

- Finally, student level error standard was use to project back to school level error standard depending on school size:

$$AssessmentSEM_{gan} = SESD_{gan} \div \sqrt{n}\,. \tag{6}$$

The results of these interpolations was an array of 491 SEMs for each of the 27 grade/subject assessments, including on-demand writing, writing portfolio, and CTBS/5 for both the baseline years and the end-of-cycle years. Note that not all school sizes are expected to be present among Kentucky schools. In a sense, these estimate are "what if" values, with estimates available for any size from 10 to 500 based on the assumptions that (1) schools near the representative sizes are similar in student error variance, and (2) interpolation between sizes follows common assumptions about variances of means (for schools) given variances in the subjects (students) making up the means.

### Accountability Index SEMs Computations

A school's accountability index for the baseline years or for the end of any of the long-term cycles is a two-year weighted average of the assessment scores available for the grades contained within the school. Consequently, SEM in the accountability index can also be computed by appropriately weighting and summing assessment SEMs.

The general formula for calculating the variance of a weighted composite from the separate variances of the individual components of the composite is:

$$\sigma^2_{Composite} = \sum_{i=1}^{n} w_i^2 \sigma_i^2 + 2\sum_{i=1}^{n}\sum_{j=1}^{n} r_{ij} w_i w_j \sigma_i \sigma_j \tag{7}$$

A composite can be decomposed into its true and error components such that some of the variance terms refer to true score variance and some to error variance. Errors are assumed to be uncorrelated with each other or with true scores, so second term components drop out with respect to error variance terms (i.e., when $r_{ij} = 0$). The resulting formula for an accountability SEM becomes:

$$AccountabilitySEM_{sc} = \sqrt{\sum w_{ac}^2 \ SEM_{as}^2}\ , \tag{8}$$

for any given combination of school size (*s*) and configuration (*c*), where the summation is over all assessments (*a*). Table A-6 in the Technical Appendix presents the assessment weights. Note that, except for the K-to-12 configuration, some assessment weights are 0.

With 14 grade configurations and 491 school sizes, 6,784 accountability SEMs were computed for the baseline years and another 6,784 accountability SEMs were computed for the end-of-cycle years. As expected, SEMs vary by both the average number of students in a grade and the number of grades in a school. They range from approximately .5 for schools with large total populations to approximately 2.5 for schools with small total populations. Note that these SEMs are "theoretical." There are not 6,784 schools in Kentucky, so most of the size-by-configuration cells in the matrix are not applicable to any particular school. Like the assessment SEMs, these accountability SEMs are "what if" values applicable given the same assumptions indicated for the assessment SEMs.

### Classification SEM Computations

The above procedures provide "look-up" tables for the various grade-configuration-by-school-size combinations for 2002/2003 and for 2003/2004 accountability SEMs. The next step is the computation of classification SEMs using the tabled values. At this point in the procedure, computational process requirements exceed the "what if" approach used for assessment and accountability SEMs. There are simply too many potential combinations of classification difference scores and accountability SEMs to create look-up tables, particularly since schools may have changed sizes between the baseline and end of cycle years.[2] Therefore, each school in the analysis is treated individually in the computation of classification SEMs.

*True Target Indexez*

Classification SEMs are a weighted function of the error variance in the baseline accountability index and error variance in the end-of-cycle accountability index where the weighting is based on the weighting use to calculate the classification index (i.e., the difference between end-of-cycle index and target). In order to calculate the classification SEM, the formula for the true target classification index is needed. Note that the true target index is not shown on the School Growth Chart or used to classify schools. On the other hand, SEM is a statistic about true scores. Therefore, the true target computations are required. Once again, there are two computations, one for the goal line and one for the assistance line. In addition, computation of the true target indexes themselves will be required in a later step of the overall process for calculating classification accuracy.

*True Goal Target*

The true goal target lies on the line connecting the baseline index (BI) in the year 2000 to the constant value of 100 in 2014. The slope of the line is:

$$Goalslope = (100 - BI) \div (2014 - 2000). \tag{9}$$

Therefore, the true target goal at the end of any cycle, where cycles (*C*) begins with Cycle 1 in 2002 is:

$$TG_c = BI + 2C((100 - BI) \div 14) = BI(1 - (2C \div 14)) + (200 \div 14)C, \tag{10}$$

which can be interpreted as a weighted function of the baseline accountability index plus a constant.

*True Assistance Target*

The assistance target for Cycle 1 ending in 2002 is simply the baseline index. For cycles 2 through 7, the true assistance line begins at the value of the baseline index plotted at 2002 and ends at 80 in 2014. The slope of this line is:

$$Asstslope = (80 - BI) \div (2014 - 2002). \tag{11}$$

Therefore, the true assistance target at the end of any cycle, where cycles (*C*) begin with Cycle 2 in 2004 is:

$$TA_c = BI + 2(C-1)((80 - BI) \div 12) = BI(1 - (2(C-1) \div 12)) + (160 \div 12)(C-1), \tag{12}$$

that can also be interpreted as a weighted function of the baseline accountability index plus a constant.

[2] If schools change configurations, special index computation rules apply frequently involving use of district level scores. These types of schools have been excluded.

*Classification and Classification SEM*

Ignoring for the moment the baseline safety net, school classification is based on the difference between a school's targets (goal and assistance) and its obtained scores:

- Positive differences from the goal indicate membership in the Meeting Goal category.
- Negative differences from the Assistance target indicate membership in the Assistance category.
- Negative differences from the goal coupled with positive differences from the Assistance target indicate membership in the Progressing category.

Calculation of classification SEMs requires only straightforward application of the formula for variance of a weighted composite, recognizing that the error variance terms are assumed to be uncorrelated. Therefore, classification accuracy for Meeting Goal versus the two lower categories is a function of error variance in the difference between $TG_c$ and the end of cycle index ($AI_c$):

$$ClassificationSEM_G = \sqrt{SEM_{AI}^2 + \left(1 - \left(2C \div 14\right)\right)^2 \times SEM_{BI}^2}\ , \tag{13}$$

where references to school size and configuration for SEMs are assumed, but not shown, and the subscript *G* refers to errors of measurement around the goal line.

In 2014 (the seventh cycle) the target for all schools is fixed at 100 and the weight for the error term reduces to 0. Error in the index goal decreases from its initial level in 2002 until it is 0 in 2014.

Classification accuracy for Assistance versus the upper categories is a function of error variance in the difference between $TA_c$ and the end of cycle Index ($AI_c$). Under the rules for computing the assistance target, in Cycle 1 the target equals the baseline accountability index. Therefore, the classification SEMs can be estimated as:

$$ClassificationSEM_A = \sqrt{SEM_{AI}^2 + SEM_{BI}^2}\ , \tag{14}$$

where reference to school size and configuration for SEMs are assumed, but not shown, and the subscript *A* refers to errors measurement associated with application of the assistance line.

For the remaining cycles, the computation incorporates a weight on the baseline error term:

$$ClassificationSEM_A = \sqrt{SEM_{AI}^2 + \left(1 - \left(2\left(C-1\right) \div 12\right)\right)^2 \times SEM_{BI}^2}\ , \tag{15}$$

where references to school size and configuration for SEMs are assumed, but not shown, and the subscript *A* refers to errors of measurement associated with application of the Assistance line.

Note that, in 2014 (the seventh cycle) the Assistance target for all schools is fixed at 80 and the weight of the Assistance error term reduces to 0.

## Shift from Standard Error to Probability Matrices

At this point, for each school eligible for the analysis, we have computed SEM for the difference scores (one for goal and one for assistance) used to classify schools. Standard errors of measurement is an index of the likely variation of observed scores around any given true score. In other words, SEM is the expected distribution of observed scores conditional on true score. Because of the effect of size and configuration on error, difference SEMs are computed for different combinations of school size and grade configuration. The same classification SEM will be computed for all schools with the same size and configuration; however, schools with the same size and configurations cannot be expected to have the same true classification difference. While individual schools have become our vehicle for determining the set of sizes and configurations for computing classification SEMs, the SEMs produced are not particularly meaningful to the individual schools because their true classification differences are unknown. Far more useful at the individual school level is the estimate of the variation in true classification differences that is expected given any particular observed classification difference. Figure 2, presented in the overview, shows the schema for making the translation. The approach uses discrete score ranges to simplify calculations. A matrix of probabilities is created for various ranges of observed scores, given set ranges for true scores. Another matrix of probabilities for various ranges of true scores, given fixed ranges for observed scores, is then created using Bayes' Theorem and estimates of true difference probabilities.

### *Creating Probability Matrix of Observed Differences Given Possible True Differences*

The matrices concern difference scores with 0 being the critical decision point, making 0 one of the required interval boundaries. After examining the range of differences between observed index scores and target index scores for goal and for assistance classification decisions, the range of differences was divided into 54 intervals. These intervals included (a) all scores less than –13, (b) all greater than +13, plus (c) the remaining 52 intervals between –13 and +13 with the width of each interval equal to .5. These same score intervals were also used for possible observed scores. For any cell in the resulting matrix, SEM values were used to calculate the probability of the identified observed difference, given the identified true difference. Calculations are based on the standard assumption that errors around any given true difference are normally distributed with standard deviation equal to the SEM.

### *Estimating of True Index Variance*

An observed assessment score is the result of a "true" score and measurement error. Likewise, variance in observed scores is a function of variance in true scores and variance in error. Since $(SEM)^2$ is an estimate of error variance, estimates of true variance are calculated by subtracting error variance from total score variance. Since the magnitude of error variance is likely a function of school size and school configuration, we assume total variance is as well. Therefore, we investigated variance in observed classification difference scores by school size and configuration.

*Calculating Total Variance in Classification Difference Scores and School Size*

In order to calculate score variance, multiple observations must be available. To create these multiple observations, schools were grouped by rounding their sizes for 2003-2004 to the nearest 25 for schools up to 300. Above 300 students per grade, schools were categorized as either 350 or 450 students per grade. Variance in classification differences for both goal and assistance targets were calculated for each of these groups. The results are plotted in Figure 3 and 4. Each figure also displays the fit of a power function to school size. The fit is very close in both cases as noted by the $R^2$ of .84 and .88 for the goal and assistance classification difference standard deviations, respectively. Given the strength of the relationship between size and variance in observed classification difference scores, using these size categories to estimate variance estimates is warranted. In contrast, there was no discernable pattern to the classification standard deviations for the different configurations and several of the configurations contains so few schools that estimated standard deviations were either not possible or potentially unstable.

*Estimating Distributions of True Differences*

Having established estimates of total variance that can be applied to schools of any given size and having calculated error variance estimates for each school given its size and configuration, true variance estimates applicable for each school were calculated as the difference between the two. The next step was to use these true variance estimates to calculate probabilities of school true scores being in any of the scores intervals (-13 to +13). Computation of the array of true difference probabilities is based on the assumption of normally distributed scores centered on the mean of the differences with a standard deviations equal to the true variance estimation. True mean differences were estimated by observed mean differences, and like total variance, mean differences were estimated separately for school size category. Figure 5 and 6 show that strength of the relationship between size and mean difference as capture by second-degree polynomial equations. Again, use of school size to capture difference means appears appropriate. Note that these true difference probability arrays are dependent on school size and configuration, but they are not yet conditioned on school observed score. That is the next step.

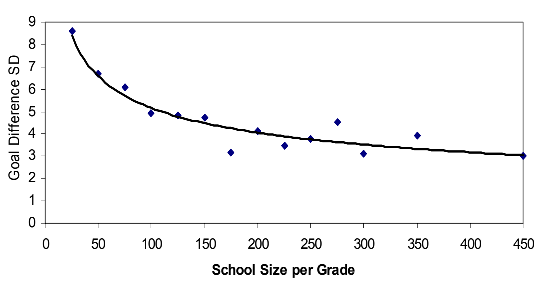



Figure 3. Classification difference standard deviations for Meeting Goal by school size.

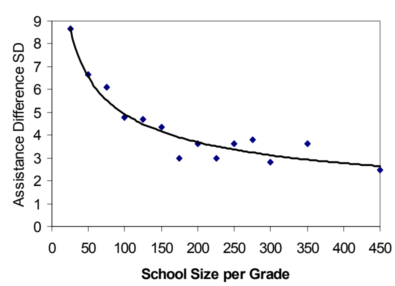



Figure 4. Classification difference standard deviations for Assistance by school size.

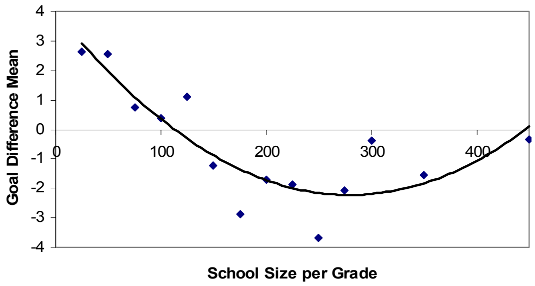



Figure 5. Classification difference means for meeting goal by school size.

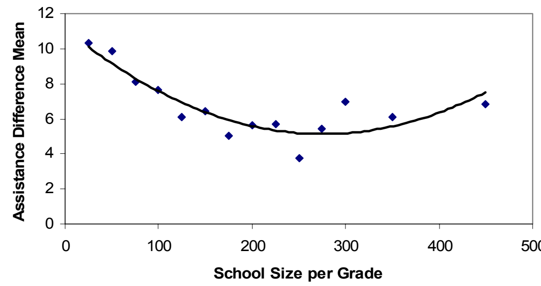



Figure 6. Classification difference means for assistance by school size.

*Creating Matrix of Probabilities of True Differences Given Observed Difference*

Once again, for each school, two matrices of probabilities of observed classification differences given true differences were calculated, one for goal decisions and one for assistance decisions. Likewise two arrays of probabilities of true differences are created for each school. For a given observed difference the probability of classification true difference in a given interval is:

$$P(True_i|Obs_j) = \frac{P(Obs_j|True_i)P(True_i)}{P(Obs_j|True_1)P(True_1)+P(Obs_j|True_2)P(True_2)+P(Obs_j|True_3)P(True_3)+\ldots+P(Obs_j|True_k)P(True_k)}, \quad (16)$$

where $Obs_j$ = observed difference represented by interval $j$, with $k$ possible difference intervals, and $True_i$ = true difference represented by interval i, with $k$ intervals represented in the probability matrix.

Bayes' transformation was applied to the data for each school. The result was a matrix of probabilities of each of the 54 true score intervals being in any of the 54 observed difference intervals, with separate matrices for meeting goal and for needing assistance. Any given school had only one observed goal difference and one observed assistance difference; therefore, only one column of either school-specific Prob(True|Observed) matrix was relevant.[3] For each school, the observed column containing the school's observed difference was identified for both the goal and assistance matrices. Appropriate summation of cell probabilities above and below zero (described earlier) provide estimates of the probabilities of the school having a true classification in Meeting Goal and in Assistance. From these two estimates, an estimate of the probability of the school having a true classification in Progressing was computed.

## Summarize Probabilities Across Schools

The final step was to summarize probabilities across school by computing mean probabilities of each of the three classifications (Meeting Goal, Progressing, and Assistance) for the observed classifications of schools (Figure 1), for the classification of school if no safety

[3] "School-specific" is not exactly correct. All schools of a given grade configuration whose sizes were identical in the base years and in the final years will have the same probability matrix.

were applied (Figure 2), and for the joint categorization that results from considering classification with and without the safety net (Figure 3).

## References

Hoffman, R. G., & Wise, L. L. (1999). *Establishing the Reliability of Student Level Classifications: Analytic Plan and Demonstration.* (HumRRO Report FR-WATSD-99-34). Alexandria, VA: Human Resources Research Organization.

Hoffman, R. G. & Wise, L. L. (2000a). *Establishing the reliability of student proficiency classifications: The accuracy of observed classifications.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, April, 2000

Hoffman, R. G., & Wise, L. L. (2000b). *School classification accuracy final analysis plan for the Commonwealth Accountability Testing System* (FR-00-26). Alexandria, VA: Human Resources Research Organization.

Hoffman, R. G. & Wise, L. L. (2001). *The Accuracy of School Classifications for the interim accountability cycle of the Kentucky Commonwealth Accountability Testing System* (FR-01-26). Alexandria, VA: Human Resources Research Organization

# Appendix
# Technical Documentations

## Generalizability Models

Standard errors of measure of the various components of the accountability model are estimated by Generalizability analyses of students' NAPD scores. Given that school index scores span two years, the basic model is one in which pupils are nested within forms, years, and schools, and forms are nested within years and are crossed with schools. For writing and for CTBS/5 forms are not a consideration, so the Generalizability model is reduced to one in which pupils are nested within schools and years.

Figure A-1 presents the four-facet design for the Kentucky Core Content Tests. Tables A-1, 2 and 3 presents the calculations using Brennan's (1981) notation and algorithms for generating sums of squares and variance components. For each of the grade/subject combinations, the six sources of variance in schools' two-year academic index averages include: (1) school, (2) year, (3) school by year, (4) form within year, (5) school by form within year, and (6) pupil within form within school by form. The order of the nesting terms in the last source of variance is a little ambiguous in its wording since pupils are nested within forms, within schools, and within years. However, for derivation of the error components, the expressed order of the nested does not matter, as long as the nesting is captured.

*Random, fixed, or sampled from a finite universe*

Generalizability theory explicitly considers the universe to which observed score are interpretable. Typically, the items that make up a particular test are only viewed as samples of an infinite array of similar items. Being sampled from an infinite domain, test items are typically considered "random." On the other hand, some facets may cover the intended universe to which scores are intended to generalize. Year, for example, could be considered fixed because the universe of generalization is two years and both years are sampled. On the other hand, year could be considered as sampled from a finite universe. The logic is this: The school academic index, while directly interpretable as the average of students' achievement, is being used to make inferences about the instructional programs of those schools. An accountability cycle is four years long. Changes in instruction that occur in any of those four years could impact students' achievement in the final two years. Thus, the universe of generalization could be viewed as instructional change that occurred in any of the four years of the cycle. Only two of the four years are assessed, however. Other than being illustrative of sampling within a fixed domain, we are making no strong argument that the present data be treated with years being samples of a fixed four-year domain. Instead, we are suggesting that years be considered fixed. Forms and pupils are assumed to be randomly sampled from a infinite domain. Table A-3 indicates that the value of for two sources of variance (year and school x year) reduce to zero when years are considered fixed.

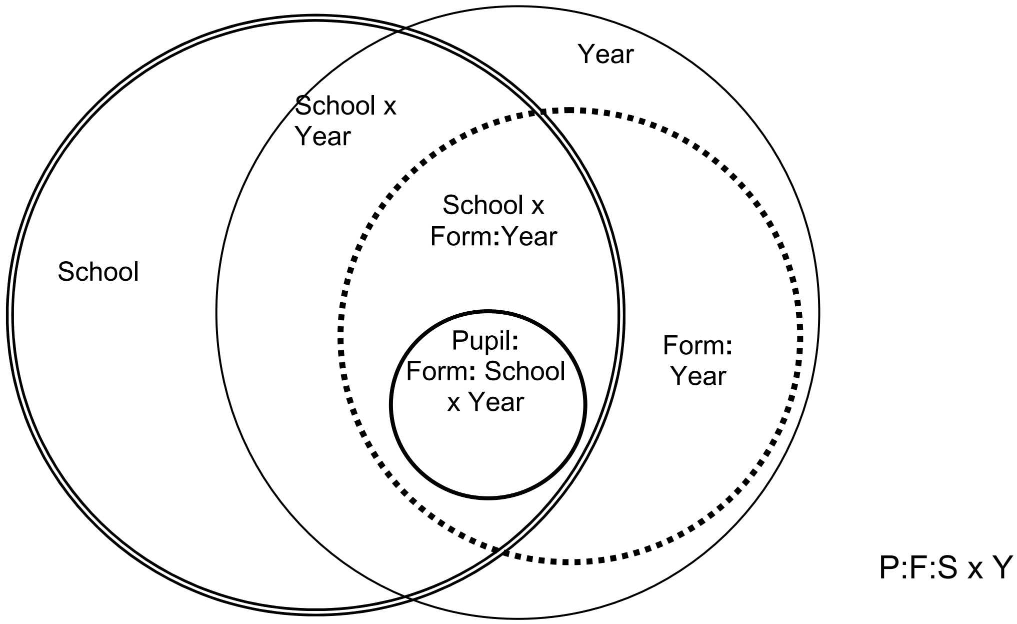



Figure A-1. Generalizability theory design representing Kentucky Core Content Test two-year accountability cycle.

Table A-1
Estimating Variance Components for Pupil: School Year Form Generalizability Theory Design – Random Effects Estimates

| Effect | df | Means | SS |
|---|---|---|---|
| School (s) | $n_s$ - 1 | $\bar{X}_s = \frac{1}{n_y n_f n_p} \sum_y \sum_f \sum_p X_{syfp}$ | $n_f\, n_y\, n_p\, \Sigma\, \bar{X}_s^2$ - $n_s\, n_y\, n_f\, n_p\ \bar{X}^2$ |
| Year (y) | $n_y$ - 1 | $\bar{X}_y = \frac{1}{n_s n_f n_p} \sum_s \sum_f \sum_p X_{syfp}$ | $n_s\, n_f\, n_p\, \Sigma\, \bar{X}_y^2$ - $n_s\, n_y\, n_f\, n_p\ \bar{X}^2$ |
| School x Year | $(n_s - 1)(n_y - 1)$ | $\bar{X}_{sy} = \frac{1}{n_f n_p} \sum_f \sum_p X_{syfp}$ | $n_f\, n_p\, \Sigma\Sigma\, \bar{X}_{sy}^2$ - $n_f\, n_y\, n_p\, \Sigma\, \bar{X}_s^2$ - $n_s\, n_f\, n_p\, \Sigma\, \bar{X}_y^2 + n_s\, n_y\, n_f\, n_p\ \bar{X}^2$ |
| Form:Year (f:y) | $n_y(n_f - 1)$ | $\bar{X}_{f:y} = \frac{1}{n_s n_p} \sum_s \sum_p X_{syfp}$ | $n_s\, n_p\, \Sigma\Sigma\, \bar{X}_{yf}^2$ - $n_s\, n_f\, n_p\, \Sigma\, \bar{X}_y^2$ |
| School x Form : Year (sf:y) | $n_y(n_s - 1)(n_f - 1)$ | $\bar{X}_{sf:y} = \frac{1}{n_p} \sum_p X_{syfp}$ | $n_p\, \Sigma\Sigma\Sigma\, \bar{X}_{syf}^2$ - $n_f\, n_p\, \Sigma\Sigma\, \bar{X}_{sy}^2$ - $n_s\, n_p\, \Sigma\Sigma\, \bar{X}_{yf}^2 + n_s\, n_f\, n_p\ \bar{X}_y^2$ |
| Pupil: School Year Form (p:sfy) | $n_y n_s n_f (n_p - 1)$ | na | $\Sigma\Sigma\Sigma\Sigma\, X_{psyf}^2$ - $n_p\, \Sigma\Sigma\Sigma\, \bar{X}_{syf}^2$ |
| Total | $n_s\, n_y\, n_f\, n_p$-1 | $\bar{X} = \frac{1}{n_s n_y n_f n_p} \sum_s \sum_y \sum_f \sum_p X_{syfp}$ | $\Sigma\Sigma\Sigma\Sigma\, X_{psyf}^2$ - $n_s\, n_y\, n_f\, n_p\, \bar{X}^2$ |

Table A-2
Estimating Variance Components for Pupil: School Year Form Generalizability Theory Design – G-Study Estimates

| Effect ($\alpha$) | Estimated $\sigma^2$ –Random Effects Model | Estimated $\sigma^2(\alpha\|M)$ -- Mixed Models (N = Universe size) | |
|---|---|---|---|
| | | Basic Mixed Model | Year Fixed |
| School (s) | $\frac{[MS(s) - MS(sy)]}{n_y n_f n_p}$ | $\hat{\sigma}^2_s + \frac{\hat{\sigma}^2_{sy}}{N_y} + \frac{\hat{\sigma}^2_{sf:y}}{N_f N_y} + \frac{\hat{\sigma}^2_{p:f:sy}}{N_f N_y N_p}$ | $\hat{\sigma}^2_s + \frac{\hat{\sigma}^2_{sy}}{N_y}$ |
| Year (y) | $\frac{[MS(y) - MS(sy) - MS(fy) + MS(sfy)]}{n_s n_f n_p}$ | $\hat{\sigma}^2_y + \frac{\hat{\sigma}^2_{sy}}{N_s} + \frac{\hat{\sigma}^2_{f:y}}{N_f} + \frac{\hat{\sigma}^2_{sf:y}}{N_s N_f} + \frac{\hat{\sigma}^2_{p:f:sy}}{N_s N_f N_p}$ | $\hat{\sigma}^2_y$ |
| School x Year | $\frac{[MS(sy) - MS(sfy)]}{n_f\, n_p}$ | $\hat{\sigma}^2_{sy} + \frac{\hat{\sigma}^2_{sf:y}}{N_f} + \frac{\hat{\sigma}^2_{p:f:sy}}{N_f N_p}$ | $\hat{\sigma}^2_{sy}$ |
| Form**:**Year (f:y) | $\frac{[MS(fy) - MS(sfy)]}{n_s\, n_p}$ | $\hat{\sigma}^2_{f:y} + \frac{\hat{\sigma}^2_{sf:y}}{N_s} + \frac{\hat{\sigma}^2_{p:f:sy}}{N_s N_p}$ | $\hat{\sigma}^2_{f:y}$ |
| School x Form : Year (sf:y) | $\frac{[MS(sfy) - MS(syfp)]}{n_p}$ | $\hat{\sigma}^2_{f:sy} + \frac{\hat{\sigma}^2_{p:f:sy}}{N_p}$ | $\hat{\sigma}^2_{f:sy}$ |
| Pupil: School Year Form (p:sfy) | MS(syfp) | $\hat{\sigma}^2_{p:f:sy}$ | $\hat{\sigma}^2_{p:f:sy}$ |

Table A-3
Estimating Variance Components for Pupil: School Year Form Generalizability Theory Design – D-study Estimates

| Effect ($\alpha$) | D-study error component | Use term in | |
|---|---|---|---|
| | | Absolute error estimate | Relative error estimate |
| School (s) | $\hat{\sigma}^2_s + \frac{\hat{\sigma}^2_{sy}}{N_y}$ | | |
| Year (y) | $[\hat{\sigma}^2_y \;/ N_y]\;[1 - \frac{n_y}{N_y}] = \mathbf{0}$ | (X) | |
| School x Year | $[\hat{\sigma}^2_{sy} \;/N_y] \times [1 - \frac{n_y}{N_y}] = \mathbf{0}$ | (X) | (X) |
| Form**:**Year (f:y) | $\hat{\sigma}^2_{f:y} \;/ N_y N_f$ | X | |
| School x Form : Year (sf:y) | $\hat{\sigma}^2_{f:sy} \;/ N_y N_f$ | X | X |
| Pupil: School Year Form (p:sfy) | $\hat{\sigma}^2_{p:f:sy} / N_y N_f\, n_p$ | X | X |

Note that current literature is mixed on whether pupils should be considered fixed, random, or sampled from a fixed domain (Cronbach, Linn, Brennan, & Haertel, 1997; Hambleton, Jaeger, Koretz, Linn, Millman, & Phillips, 1996; Yen, 1997). Persistent criticisms of Kentucky's accountability model that cohort-to-cohort variation in student proficiency is unfair (Hoffman, 1998) makes treating students as fixed unwise. Yen uses two different approaches, one for which students are random, and a second for which students are treated as samples of a finite domain with that domain being defined as the total school population from which the tested students are taken. Yen's second approach does not fit Kentucky's two year cycle very well, particularly since we know the transience among students is perceived to be a significant issue for some districts (Thacker, Koger, Hoffman, and Koger, 2000) and is indeed related to school scores (Medsker, 1998). Therefore, we have chosen to treat students as random, i.e., sampled from an infinite universe. (Note also that in Yen's second approach, she adds a term for measurement error at the person level. That term is mathematically eliminated when students are treated as random.)

Yen (1997) also discussed potential modification to the forms by schools interaction given that forms are intended to target slightly different content. She concludes that since there is no way to directly test differences in targets (forms and students are confounded), the straightforward approach, as presented in Tables 2 – 4, is more acceptable with a caveat that it may overestimate standard error.

*Absolute and relative error*

Generalizability theory considers two kinds of error: absolute and relative. Absolute error is appropriate to consider when the objects of measure (schools in our case) are being assessed against a standard that generalizes beyond any of the particular instances of the various facets of measurement (e.g., different forms, different years, different pupils). Relative error, on the other hand, is appropriate when schools are being compared to each other and have been subject to the same measurement processes (same forms, same years). Table A-3 indicates which variance components enter each type of error estimate. With years treated as fixed, three error components (form within year, school by form within year, and pupil within form within school by form) are summed to estimate absolute error. Only the later two components (school by form within year, and pupil within form within school by form) are summed to estimate error variance for the relative model. Because schools are being assessed against a standard, rather than by relative standing among other schools, absolute error is the appropriate estimate to use in computing CATS classification accuracy.

*Special Considerations for Writing Assessments*

Each student completes one on-demand writing prompt, and it is chosen by the student from a pair of alternatives. Six pairs of writing prompts constitute six forms for on-demand writing. From past analysis (Hoffman, Koger, & Awbrey,1997), we know that means for different writing prompts vary greatly for prompts within a form as well as for prompts from different forms. The variation in means leads to the conclusion that each prompt should be treated as a separate "form" using the same Generalizability analysis design described above. As far as the self-selection factor in concerned, we see no option other than considering it one of the random factors affecting prompt (i.e., item) sampling.

Portfolios, however, are (in theory[4]) unique to each individual student. "Forms" as a theoretical facet for portfolios is confounded with students.[5] Therefore, school-level error variance for portfolios will be assessed using a Generalizability design similar to the one presented above, but without form as a facet. That is, pupils are nested within the intersection of schools and years. Formulas for this three facet (pupils:schools x years) are available in Brennan (1981), designated as i:(p x h) in his notation.

*CTBS/5*

CTBS/5 scores also do not include separate forms at any one of the grade levels in which it is administered. Therefore the same Generalizability model applied to writing portfolios is applied to CTBS/5 scores.

[4] Some schools do tend to structure common activities and present selected topics for students to create portfolio entries.
[5] Again, this is an oversimplication. Anecdotally, some schools reportedly have been doing a better job than others of structuring portfolio activities that facilitate higher quality writing. "Item sampling," therefore, may be confounded with schools. In this unusual case, schools become both the object of measurement and an instrument, or facet, of measurement.

Table A-4
Variance Components for Each Grade/Subject By School size Configuration for baseline 1999-2000

| | | | | |
|---|---|---|---|---|
| rd = Reading<br>sc = Science<br>wo = Writing Prompt<br>wp = Writing Portfolio<br>ah = Arts & Humanities<br>ma = Mathematics<br>pl = PL/VS<br>ss = Social Studies | Lg = Large School<br>Md = Medium School<br>Sm = Small School | NS = Number of Schools<br>NP = Number of Pupils<br>NF = Number of Forms<br>NY = Number of Years | Ab, Err = Absolute Error Variance<br>Rel. Error = Relative Error Variance<br>Tot Var. = Total Variance | Ab. Gen. = Absolute Generalizability<br>Rel. Gen. = Relative Generalizability |

| Grade | Subject | School Size | NS | NP | NY | NF | Absol. Err. | Rel. Err. | Total Var. | Absol. Gen. | Rel. Gen. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 3 | ct | lg | 66 | 96 | 2 | . | 11.935 | 11.935 | 281.277 | 0.958 | 0.958 |
| 3 | ct | md | 35 | 60 | 2 | . | 18.568 | 18.568 | 406.480 | 0.954 | 0.954 |
| 3 | ct | sm | 49 | 24 | 2 | . | 48.407 | 48.407 | 275.457 | 0.824 | 0.824 |
| 4 | rd | lg | 36 | 16 | 2 | 6 | 6.208 | 5.995 | 101.721 | 0.939 | 0.941 |
| 4 | rd | md | 55 | 10 | 2 | 6 | 8.106 | 8.028 | 140.524 | 0.942 | 0.943 |
| 4 | rd | sm | 44 | 4 | 2 | 6 | 22.325 | 21.798 | 75.395 | 0.704 | 0.711 |
| 4 | sc | lg | 36 | 16 | 2 | 6 | 6.119 | 6.119 | 110.375 | 0.945 | 0.945 |
| 4 | sc | md | 55 | 10 | 2 | 6 | 7.821 | 7.821 | 182.917 | 0.957 | 0.957 |
| 4 | sc | sm | 44 | 4 | 2 | 6 | 18.186 | 17.839 | 108.250 | 0.832 | 0.835 |
| 4 | wod | lg | 35 | 16 | 2 | 6 | 5.651 | 5.512 | 44.072 | 0.872 | 0.875 |
| 4 | wod | md | 54 | 10 | 2 | 6 | 7.972 | 7.896 | 52.788 | 0.849 | 0.850 |
| 4 | wod | sm | 42 | 4 | 2 | 6 | 15.894 | 15.867 | 47.132 | 0.663 | 0.663 |
| 4 | wp | lg | 54 | 96 | 2 | . | 4.048 | 4.048 | 147.104 | 0.972 | 0.972 |
| 4 | wp | md | 29 | 60 | 2 | . | 6.090 | 6.090 | 199.888 | 0.970 | 0.970 |
| 4 | wp | sm | 51 | 24 | 2 | . | 17.683 | 17.683 | 227.601 | 0.922 | 0.922 |
| 5 | ah | lg | 28 | 8 | 2 | 12 | 8.067 | 7.939 | 143.255 | 0.944 | 0.945 |
| 5 | ah | md | 39 | 5 | 2 | 12 | 10.796 | 10.459 | 119.436 | 0.910 | 0.912 |
| 5 | ah | sm | 28 | 2 | 2 | 12 | 22.270 | 22.175 | 85.604 | 0.740 | 0.741 |
| 5 | ma | lg | 33 | 16 | 2 | 6 | 7.364 | 7.186 | 200.391 | 0.963 | 0.964 |
| 5 | ma | md | 57 | 10 | 2 | 6 | 9.426 | 9.426 | 178.516 | 0.947 | 0.947 |
| 5 | ma | sm | 39 | 4 | 2 | 6 | 22.213 | 22.213 | 145.874 | 0.848 | 0.848 |
| 5 | pl | lg | 28 | 8 | 2 | 12 | 8.868 | 8.632 | 142.296 | 0.938 | 0.939 |
| 5 | pl | md | 38 | 5 | 2 | 12 | 12.913 | 12.737 | 131.288 | 0.902 | 0.903 |
| 5 | pl | sm | 28 | 2 | 2 | 12 | 31.440 | 31.440 | 156.133 | 0.799 | 0.799 |
| 5 | ss | lg | 32 | 16 | 2 | 6 | 8.144 | 8.144 | 229.568 | 0.965 | 0.965 |
| 5 | ss | md | 57 | 10 | 2 | 6 | 12.491 | 12.312 | 199.534 | 0.937 | 0.938 |
| 5 | ss | sm | 39 | 4 | 2 | 6 | 27.197 | 26.125 | 199.494 | 0.864 | 0.869 |
| 6 | ct | lg | 36 | 180 | 2 | . | 6.494 | 6.494 | 159.455 | 0.959 | 0.959 |
| 6 | ct | md | 42 | 60 | 2 | . | 18.344 | 18.344 | 335.471 | 0.945 | 0.945 |
| 6 | ct | sm | 41 | 24 | 2 | . | 49.311 | 49.311 | 181.653 | 0.729 | 0.729 |
| 7 | rd | lg | 41 | 40 | 2 | 6 | 2.293 | 2.205 | 122.577 | 0.981 | 0.982 |
| 7 | rd | md | 22 | 20 | 2 | 6 | 4.428 | 4.230 | 49.878 | 0.911 | 0.915 |
| 7 | rd | sm | 28 | 6 | 2 | 6 | 12.799 | 12.799 | 107.816 | 0.881 | 0.881 |
| 7 | sc | lg | 41 | 40 | 2 | 6 | 3.591 | 3.571 | 173.255 | 0.979 | 0.979 |
| 7 | sc | md | 22 | 20 | 2 | 6 | 7.215 | 7.215 | 80.072 | 0.910 | 0.910 |
| 7 | sc | sm | 28 | 6 | 2 | 6 | 15.238 | 14.478 | 187.607 | 0.919 | 0.923 |

Table A-4
Variance Components for Each Grade/Subject By School size Configuration for baseline 1999-2000

| | | | | |
|---|---|---|---|---|
| rd = Reading<br>sc = Science<br>wo = Writing Prompt<br>wp = Writing Portfolio<br>ah = Arts & Humanities<br>ma = Mathematics<br>pl = PL/VS<br>ss = Social Studies | Lg = Large School<br>Md = Medium School<br>Sm = Small School | NS = Number of Schools<br>NP = Number of Pupils<br>NF = Number of Forms<br>NY = Number of Years | Ab, Err = Absolute Error Variance<br>Rel. Error = Relative Error Variance<br>Tot Var. = Total Variance | Ab. Gen. = Absolute Generalizability<br>Rel. Gen. = Relative Generalizability |

| Grade | Subject | School Size | NS | NP | NY | NF | Absol. Err. | Rel. Err. | Total Var. | Absol. Gen. | Rel. Gen. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 7 | wod | lg | 41 | 40 | 2 | 6 | 2.510 | 2.260 | 64.101 | 0.961 | 0.965 |
| 7 | wod | md | 22 | 20 | 2 | 6 | 4.330 | 4.249 | 30.428 | 0.858 | 0.860 |
| 7 | wod | sm | 27 | 6 | 2 | 6 | 11.789 | 11.789 | 75.778 | 0.844 | 0.844 |
| 7 | wp | lg | 48 | 240 | 2 | . | 1.733 | 1.733 | 148.413 | 0.988 | 0.988 |
| 7 | wp | md | 27 | 120 | 2 | . | 3.700 | 3.700 | 69.190 | 0.947 | 0.947 |
| 7 | wp | sm | 36 | 36 | 2 | . | 12.672 | 12.672 | 120.415 | 0.895 | 0.895 |
| 8 | ah | lg | 29 | 20 | 2 | 12 | 3.241 | 3.208 | 126.937 | 0.974 | 0.975 |
| 8 | ah | md | 26 | 10 | 2 | 12 | 6.147 | 6.061 | 106.441 | 0.942 | 0.943 |
| 8 | ah | sm | 21 | 3 | 2 | 12 | 17.900 | 17.649 | 270.439 | 0.934 | 0.935 |
| 8 | ma | lg | 40 | 40 | 2 | 6 | 2.484 | 2.446 | 128.201 | 0.981 | 0.981 |
| 8 | ma | md | 27 | 20 | 2 | 6 | 4.868 | 4.781 | 79.019 | 0.938 | 0.939 |
| 8 | ma | sm | 26 | 6 | 2 | 6 | 13.543 | 13.543 | 345.025 | 0.961 | 0.961 |
| 8 | pl | lg | 30 | 20 | 2 | 12 | 3.398 | 3.356 | 108.562 | 0.969 | 0.969 |
| 8 | pl | md | 26 | 10 | 2 | 12 | 7.395 | 7.356 | 104.654 | 0.929 | 0.930 |
| 8 | pl | sm | 20 | 3 | 2 | 12 | 22.297 | 22.297 | 257.481 | 0.913 | 0.913 |
| 8 | ss | lg | 41 | 40 | 2 | 6 | 3.185 | 3.185 | 108.854 | 0.971 | 0.971 |
| 8 | ss | md | 27 | 20 | 2 | 6 | 5.375 | 5.191 | 109.991 | 0.951 | 0.953 |
| 8 | ss | sm | 26 | 6 | 2 | 6 | 12.817 | 12.455 | 273.806 | 0.953 | 0.955 |
| 9 | ct | lg | 46 | 312 | 2 | . | 4.143 | 4.143 | 327.514 | 0.987 | 0.987 |
| 9 | ct | md | 36 | 168 | 2 | . | 7.733 | 7.733 | 236.606 | 0.967 | 0.967 |
| 9 | ct | sm | 36 | 24 | 2 | . | 53.091 | 53.091 | 305.827 | 0.826 | 0.826 |
| 10 | pl | lg | 47 | 20 | 2 | 12 | 3.276 | 3.190 | 106.937 | 0.969 | 0.970 |
| 10 | pl | md | 29 | 14 | 2 | 12 | 5.655 | 5.495 | 65.694 | 0.914 | 0.916 |
| 10 | pl | sm | 26 | 5 | 2 | 12 | 12.844 | 12.844 | 65.829 | 0.805 | 0.805 |
| 10 | rd | lg | 56 | 40 | 2 | 6 | 2.392 | 2.338 | 102.136 | 0.977 | 0.977 |
| 10 | rd | md | 39 | 28 | 2 | 6 | 3.839 | 3.748 | 61.919 | 0.938 | 0.939 |
| 10 | rd | sm | 29 | 10 | 2 | 6 | 8.099 | 8.099 | 65.020 | 0.875 | 0.875 |
| 11 | ah | lg | 35 | 20 | 2 | 12 | 3.443 | 3.365 | 161.583 | 0.979 | 0.979 |
| 11 | ah | md | 24 | 14 | 2 | 12 | 4.278 | 4.218 | 101.996 | 0.958 | 0.959 |
| 11 | ah | sm | 34 | 5 | 2 | 12 | 10.347 | 10.321 | 102.731 | 0.899 | 0.900 |
| 11 | ma | lg | 40 | 40 | 2 | 6 | 3.068 | 2.840 | 168.993 | 0.982 | 0.983 |
| 11 | ma | md | 27 | 28 | 2 | 6 | 3.814 | 3.750 | 172.664 | 0.978 | 0.978 |
| 11 | ma | sm | 38 | 10 | 2 | 6 | 9.492 | 9.249 | 102.219 | 0.907 | 0.910 |
| 11 | sc | lg | 40 | 40 | 2 | 6 | 2.923 | 2.754 | 96.554 | 0.970 | 0.971 |
| 11 | sc | md | 27 | 28 | 2 | 6 | 3.194 | 2.849 | 78.908 | 0.960 | 0.964 |
| 11 | sc | sm | 38 | 10 | 2 | 6 | 7.591 | 7.519 | 74.248 | 0.898 | 0.899 |
| 11 | ss | lg | 40 | 40 | 2 | 6 | 2.352 | 2.310 | 140.559 | 0.983 | 0.984 |
| 11 | ss | md | 27 | 28 | 2 | 6 | 2.871 | 2.753 | 99.381 | 0.971 | 0.972 |

Table A-4
Variance Components for Each Grade/Subject By School size Configuration for baseline 1999-2000

| | | | | |
|---|---|---|---|---|
| rd = Reading<br>sc = Science<br>wo = Writing Prompt<br>wp = Writing Portfolio<br>ah = Arts & Humanities<br>ma = Mathematics<br>pl = PL/VS<br>ss = Social Studies | Lg = Large School<br>Md = Medium School<br>Sm = Small School | NS = Number of Schools<br>NP = Number of Pupils<br>NF = Number of Forms<br>NY = Number of Years | Ab, Err = Absolute Error Variance<br>Rel. Error = Relative Error Variance<br>Tot Var. = Total Variance | Ab. Gen. = Absolute Generalizability<br>Rel. Gen. = Relative Generalizability |

| Grade | Subject | School Size | NS | NP | NY | NF | Absol. Err. | Rel. Err. | Total Var. | Absol. Gen. | Rel. Gen. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 11 | ss | sm | 38 | 10 | 2 | 6 | 7.874 | 7.874 | 75.181 | 0.895 | 0.895 |
| 12 | wod | lg | 29 | 40 | 2 | 6 | 1.673 | 1.606 | 21.860 | 0.923 | 0.927 |
| 12 | wod | md | 29 | 28 | 2 | 6 | 2.853 | 2.636 | 37.943 | 0.925 | 0.931 |
| 12 | wod | sm | 29 | 10 | 2 | 6 | 6.263 | 6.263 | 40.971 | 0.847 | 0.847 |
| 12 | wp | lg | 36 | 240 | 2 | . | 1.991 | 1.991 | 61.669 | 0.968 | 0.968 |
| 12 | wp | md | 50 | 168 | 2 | . | 3.002 | 3.002 | 82.675 | 0.964 | 0.964 |
| 12 | wp | sm | 42 | 60 | 2 | . | 7.959 | 7.959 | 92.523 | 0.914 | 0.914 |

Table A-5
Variance Components for Each Grade/Subject By School size Configuration for End-of-Cycle 2003-2004

| | | | | |
|---|---|---|---|---|
| rd = Reading<br>sc = Science<br>wo = Writing Prompt<br>wp = Writing Portfolio<br>ah = Arts & Humanities<br>ma = Mathematics<br>pl = PL/VS<br>ss = Social Studies | Lg = Large School<br>Md = Medium School<br>Sm = Small School | NS = Number of Schools<br>NP = Number of Pupils<br>NF = Number of Forms<br>NY = Number of Years | Ab, Err = Absolute Error Variance<br>Rel. Error = Relative Error Variance<br>Tot Var. = Total Variance | Ab. Gen. = Absolute Generalizability<br>Rel. Gen. = Relative Generalizability |

| Grade | Subject | School Size | NS | NP | NY | NF | Absol. Err. | Rel. Err. | Total Var. | Absol. Gen. | Rel. Gen. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 3 | ct | lg | 45 | 96 | 2 | . | 10.004 | 10.004 | 167.919 | 0.940 | 0.940 |
| 3 | ct | md | 41 | 60 | 2 | . | 17.007 | 17.007 | 230.608 | 0.926 | 0.926 |
| 3 | ct | Sm | 43 | 24 | 2 | . | 42.384 | 42.384 | 218.827 | 0.806 | 0.806 |
| 4 | rd | lg | 33 | 16 | 2 | 6 | 4.495 | 4.495 | 88.967 | 0.949 | 0.949 |
| 4 | rd | md | 19 | 10 | 2 | 6 | 7.694 | 7.507 | 77.002 | 0.900 | 0.903 |
| 4 | rd | sm | 30 | 4 | 2 | 6 | 17.789 | 17.576 | 96.630 | 0.816 | 0.818 |
| 4 | sc | lg | 33 | 16 | 2 | 6 | 4.910 | 4.885 | 135.920 | 0.964 | 0.964 |
| 4 | sc | md | 19 | 10 | 2 | 6 | 8.049 | 7.815 | 82.432 | 0.902 | 0.905 |
| 4 | sc | sm | 30 | 4 | 2 | 6 | 16.599 | 16.555 | 145.675 | 0.884 | 0.885 |
| 4 | wd | lg | 27 | 16 | 2 | 6 | 6.728 | 4.522 | 52.716 | 0.872 | 0.914 |
| 4 | wd | md | 10 | 10 | 2 | 6 | 7.873 | 7.175 | 77.972 | 0.899 | 0.908 |
| 4 | wd | sm | 29 | 4 | 2 | 6 | 16.675 | 14.167 | 81.602 | 0.796 | 0.826 |
| 4 | wp | lg | 44 | 96 | 2 | . | 2.835 | 2.835 | 128.416 | 0.978 | 0.978 |
| 4 | wp | md | 33 | 60 | 2 | . | 4.561 | 4.561 | 154.237 | 0.970 | 0.970 |
| 4 | wp | sm | 36 | 24 | 2 | . | 13.508 | 13.508 | 195.878 | 0.931 | 0.931 |
| 5 | ah | lg | 25 | 8 | 2 | 12 | 7.662 | 7.373 | 140.308 | 0.945 | 0.947 |
| 5 | ah | md | 5 | 5 | 2 | 12 | 15.628 | 15.628 | 348.627 | 0.955 | 0.955 |
| 5 | ah | sm | 27 | 2 | 2 | 12 | 20.184 | 20.184 | 167.725 | 0.879 | 0.879 |
| 5 | ma | lg | 34 | 16 | 2 | 6 | 7.172 | 7.172 | 163.811 | 0.956 | 0.956 |
| 5 | ma | md | 17 | 10 | 2 | 6 | 10.968 | 10.968 | 168.182 | 0.935 | 0.935 |
| 5 | ma | sm | 30 | 4 | 2 | 6 | 25.342 | 24.848 | 173.636 | 0.854 | 0.858 |
| 5 | pl | lg | 25 | 8 | 2 | 12 | 6.817 | 6.604 | 107.239 | 0.936 | 0.938 |
| 5 | pl | md | 5 | 5 | 2 | 12 | 12.105 | 12.105 | 220.741 | 0.945 | 0.945 |
| 5 | pl | sm | 27 | 2 | 2 | 12 | 28.768 | 28.561 | 156.052 | 0.816 | 0.817 |
| 5 | ss | lg | 34 | 16 | 2 | 6 | 6.665 | 6.661 | 177.193 | 0.962 | 0.962 |
| 5 | ss | md | 17 | 10 | 2 | 6 | 12.202 | 12.202 | 204.863 | 0.940 | 0.940 |
| 5 | ss | sm | 30 | 4 | 2 | 6 | 26.709 | 26.709 | 158.571 | 0.832 | 0.832 |
| 6 | ct | lg | 46 | 180 | 2 | . | 6.343 | 6.343 | 267.664 | 0.976 | 0.976 |
| 6 | ct | md | 30 | 60 | 2 | . | 18.418 | 18.418 | 190.093 | 0.903 | 0.903 |
| 6 | ct | sm | 32 | 24 | 2 | . | 44.544 | 44.544 | 196.381 | 0.773 | 0.773 |
| 7 | rd | lg | 48 | 40 | 2 | 6 | 1.932 | 1.738 | 69.626 | 0.972 | 0.975 |
| 7 | rd | md | 18 | 20 | 2 | 6 | 3.696 | 8.696 | 75.975 | 0.951 | 0.951 |
| 7 | rd | sm | 29 | 6 | 2 | 6 | 11.156 | 10.735 | 140.306 | 0.920 | 0.923 |
| 7 | sc | lg | 48 | 40 | 2 | 6 | 3.366 | 3.243 | 137.861 | 0.976 | 0.976 |
| 7 | sc | md | 18 | 20 | 2 | 6 | 5.818 | 5.398 | 191.637 | 0.970 | 0.972 |
| 7 | sc | sm | 29 | 6 | 2 | 6 | 17.805 | 17.629 | 274.434 | 0.935 | 0.936 |

Table A-5
Variance Components for Each Grade/Subject By School size Configuration for End-of-Cycle 2003-2004

| | | | | |
|---|---|---|---|---|
| rd = Reading<br>sc = Science<br>wo = Writing Prompt<br>wp = Writing Portfolio<br>ah = Arts & Humanities<br>ma = Mathematics<br>pl = PL/VS<br>ss = Social Studies | Lg = Large School<br>Md = Medium School<br>Sm = Small School | NS = Number of Schools<br>NP = Number of Pupils<br>NF = Number of Forms<br>NY = Number of Years | Ab, Err = Absolute Error Variance<br>Rel. Error = Relative Error Variance<br>Tot Var. = Total Variance | Ab. Gen. = Absolute Generalizability<br>Rel. Gen. = Relative Generalizability |

| Grade | Subject | School Size | NS | NP | NY | NF | Absol. Err. | Rel. Err. | Total Var. | Absol. Gen. | Rel. Gen. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 7 | wo | lg | 55 | 16 | 2 | 6 | 4.332 | 4.131 | 50.258 | 0.914 | 0.918 |
| 7 | wo | md | 27 | 10 | 2 | 6 | 5.981 | 5.795 | 53.171 | 0.888 | 0.891 |
| 7 | wo | sm | 5 | 6 | 2 | 6 | 11.690 | 10.302 | 62.701 | 0.814 | 0.836 |
| 7 | wp | lg | 58 | 240 | 2 | . | 1.842 | 1.842 | 188.371 | 0.990 | 0.990 |
| 7 | wp | md | 24 | 120 | 2 | . | 3.654 | 3.654 | 112.742 | 0.968 | 0.968 |
| 7 | wp | sm | 38 | 36 | 2 | . | 12.586 | 12.586 | 146.029 | 0.914 | 0.914 |
| 8 | ah | lg | 43 | 20 | 2 | 12 | 3.930 | 3.719 | 147.862 | 0.973 | 0.975 |
| 8 | ah | md | 16 | 10 | 2 | 12 | 6.772 | 6.689 | 220.781 | 0.969 | 0.970 |
| 8 | ah | sm | 19 | 3 | 2 | 12 | 19.489 | 19.372 | 320.170 | 0.939 | 0.940 |
| 8 | ma | lg | 49 | 40 | 2 | 6 | 2.696 | 2.563 | 126.443 | 0.979 | 0.980 |
| 8 | ma | md | 18 | 20 | 2 | 6 | 4.975 | 4.975 | 139.513 | 0.964 | 0.964 |
| 8 | ma | sm | 35 | 6 | 2 | 6 | 14.672 | 14.672 | 259.975 | 0.944 | 0.944 |
| 8 | pl | lg | 43 | 20 | 2 | 12 | 3.711 | 3.663 | 96.383 | 0.962 | 0.962 |
| 8 | pl | md | 16 | 10 | 2 | 12 | 6.611 | 6.340 | 137.607 | 0.952 | 0.954 |
| 8 | pl | sm | 19 | 3 | 2 | 12 | 18.345 | 18.214 | 211.639 | 0.913 | 0.914 |
| 8 | ss | lg | 49 | 40 | 2 | 6 | 3.286 | 3.088 | 123.386 | 0.973 | 0.975 |
| 8 | ss | md | 18 | 20 | 2 | 6 | 4.758 | 4.758 | 133.005 | 0.964 | 0.964 |
| 8 | ss | sm | 35 | 6 | 2 | 6 | 15.580 | 14.855 | 183.452 | 0.915 | 0.919 |
| 9 | ct | lg | 80 | 240 | 2 | . | 5.397 | 5.397 | 249.632 | 0.978 | 0.978 |
| 9 | ct | md | 43 | 168 | 2 | . | 7.846 | 7.846 | 180.539 | 0.957 | 0.957 |
| 9 | ct | sm | 33 | 42 | 2 | . | 30.178 | 30.178 | 245.419 | 0.877 | 0.877 |
| 10 | pl | lg | 54 | 20 | 2 | 12 | 3.814 | 3.681 | 82.298 | 0.954 | 0.955 |
| 10 | pl | md | 17 | 14 | 2 | 12 | 5.495 | 5.344 | 79.838 | 0.931 | 0.933 |
| 10 | pl | sm | 28 | 5 | 2 | 12 | 13.546 | 13.380 | 123.802 | 0.891 | 0.892 |
| 10 | rd | lg | 59 | 40 | 2 | 6 | 3.072 | 2.677 | 126.643 | 0.976 | 0.979 |
| 10 | rd | md | 25 | 28 | 2 | 6 | 3.432 | 3.249 | 80.403 | 0.957 | 0.960 |
| 10 | rd | sm | 32 | 10 | 2 | 6 | 9.791 | 9.383 | 121.608 | 0.919 | 0.923 |
| 11 | ah | lg | 36 | 20 | 2 | 12 | 4.498 | 4.400 | 175.102 | 0.974 | 0.975 |
| 11 | ah | md | 22 | 14 | 2 | 12 | 6.245 | 6.052 | 130.195 | 0.952 | 0.954 |
| 11 | ah | sm | 30 | 5 | 2 | 12 | 14.428 | 14.251 | 164.812 | 0.912 | 0.914 |
| 11 | ma | lg | 39 | 40 | 2 | 6 | 3.473 | 3.352 | 157.860 | 0.978 | 0.979 |
| 11 | ma | md | 28 | 28 | 2 | 6 | 4.532 | 4.261 | 130.999 | 0.965 | 0.967 |
| 11 | ma | sm | 34 | 10 | 2 | 6 | 11.721 | 11.656 | 143.732 | 0.918 | 0.919 |
| 11 | sc | lg | 39 | 40 | 2 | 6 | 2.725 | 2.534 | 70.717 | 0.961 | 0.964 |
| 11 | sc | md | 28 | 28 | 2 | 6 | 3.253 | 3.052 | 64.424 | 0.950 | 0.953 |
| 11 | sc | sm | 34 | 10 | 2 | 6 | 9.634 | 9.634 | 96.570 | 0.900 | 0.900 |
| 11 | ss | lg | 39 | 40 | 2 | 6 | 3.603 | 3.255 | 136.390 | 0.974 | 0.976 |
| 11 | ss | md | 28 | 28 | 2 | 6 | 4.495 | 4.197 | 98.326 | 0.954 | 0.957 |

Table A-5
Variance Components for Each Grade/Subject By School size Configuration for End-of-Cycle 2003-2004

| rd = Reading<br>sc = Science<br>wo = Writing Prompt<br>wp = Writing Portfolio<br>ah = Arts & Humanities<br>ma = Mathematics<br>pl = PL/VS<br>ss = Social Studies | Lg =<br>Large School<br>Md =<br>Medium School<br>Sm =<br>Small School | NS =<br>Number of Schools<br>NP =<br>Number of Pupils<br>NF =<br>Number of Forms<br>NY =<br>Number of Years | Ab, Err =<br>Absolute Error Variance<br>Rel. Error =<br>Relative Error Variance<br>Tot Var. =<br>Total Variance | Ab. Gen. =<br>Absolute Generalizability<br>Rel. Gen. =<br>Relative Generalizability |
|---|---|---|---|---|

| Grade | Subject | School Size | NS | NP | NY | NF | Absol. Err. | Rel. Err. | Total Var. | Absol. Gen. | Rel. Gen. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 11 | ss | sm | 34 | 10 | 2 | 6 | 10.489 | 10.126 | 119.414 | 0.912 | 0.915 |
| 12 | wo | lg | 14 | 38 | 2 | 6 | 2.398 | 1.580 | 46.627 | 0.952 | 0.968 |
| 12 | wo | md | 12 | 26 | 2 | 6 | 3.071 | 2.855 | 51.158 | 0.940 | 0.944 |
| 12 | wo | sm | 24 | 9 | 2 | 6 | 15.875 | 7.903 | 73.014 | 0.783 | 0.892 |
| 12 | wp | lg | 39 | 123 | 2 | . | 13.848 | 13.848 | 223.017 | 0.938 | 0.938 |
| 12 | wp | md | 35 | 86 | 2 | . | 17.765 | 17.765 | 227.747 | 0.922 | 0.922 |
| 12 | wp | sm | 43 | 31 | 2 | . | 54.895 | 54.895 | 267.423 | 0.795 | 0.795 |

## Weights used in Calculating accountability index Score and accountability index SEMs

Table A-6 Weight used in Calculating accountability index Score and accountability index SEMs

| Grade | Subject | WK_5 | WK_6 | WK_8 | WK_12 | W4_5 | W4_6 | W4_8 | W6_8 | W6_12 | W7_8 | W7_9 | W7_12 | W9_12 | W10_12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 03 | ct | .050000 | .025000 | .025000 | .016667 | | | | | | | | | | |
| 04 | rd | .190000 | .190000 | .095000 | .063333 | .200000 | .190000 | .100000 | | | | | | | |
| 04 | sc | .142500 | .142500 | .071250 | .047500 | .150000 | .142500 | .075000 | | | | | | | |
| 04 | wod | .028500 | .028500 | .014250 | .009500 | .030000 | .028500 | .015000 | | | | | | | |
| 04 | wp | .114000 | .114000 | .057000 | .038000 | .120000 | .114000 | .060000 | | | | | | | |
| 05 | ah | .047500 | .047500 | .023750 | .015833 | .050000 | .047500 | .025000 | | | | | | | |
| 05 | ma | .190000 | .190000 | .095000 | .063333 | .200000 | .190000 | .100000 | | | | | | | |
| 05 | na | .047500 | .047500 | .023750 | .015833 | .050000 | .047500 | .025000 | | | | | | | |
| 05 | pl | .047500 | .047500 | .023750 | .015833 | .050000 | .047500 | .025000 | | | | | | | |
| 05 | ss | .142500 | .142500 | .071250 | .047500 | .150000 | .142500 | .075000 | | | | | | | |
| 06 | ct | | .025000 | .025000 | .016667 | | .050000 | .025000 | .050000 | .025000 | | | | | |
| 07 | rd | | | .071250 | .047500 | | | .071250 | .142500 | .071250 | .150000 | .142500 | .075000 | | |
| 07 | sc | | | .071250 | .047500 | | | .071250 | .142500 | .071250 | .150000 | .142500 | .075000 | | |
| 07 | wod | | | .014250 | .009500 | | | .014250 | .028500 | .014250 | .030000 | .028500 | .015000 | | |
| 07 | wp | | | .057000 | .038000 | | | .057000 | .114000 | .057000 | .120000 | .114000 | .060000 | | |
| 08 | ah | | | .035625 | .023750 | | | .035625 | .071250 | .035625 | .075000 | .071250 | .037500 | | |
| 08 | ma | | | .071250 | .047500 | | | .071250 | .142500 | .071250 | .150000 | .142500 | .075000 | | |
| 08 | na | | | .047500 | .031667 | | | .047500 | .095000 | .047500 | .100000 | .095000 | .050000 | | |
| 08 | pl | | | .035625 | .023750 | | | .035625 | .071250 | .035625 | .075000 | .071250 | .037500 | | |
| 08 | ss | | | .071250 | .047500 | | | .071250 | .142500 | .071250 | .150000 | .142500 | .075000 | | |
| 09 | ct | | | | .016667 | | | | | .025000 | | .050000 | .025000 | .050000 | |
| 10 | pl | | | | .023750 | | | | | .035625 | | | .035625 | .071250 | .075000 |
| 10 | rd | | | | .047500 | | | | | .071250 | | | .071250 | .142500 | .150000 |
| 11 | ah | | | | .023750 | | | | | .035625 | | | .035625 | .071250 | .075000 |
| 11 | ma | | | | .047500 | | | | | .071250 | | | .071250 | .142500 | .150000 |
| 11 | sc | | | | .047500 | | | | | .071250 | | | .071250 | .142500 | .150000 |
| 11 | ss | | | | .047500 | | | | | .071250 | | | .071250 | .142500 | .150000 |
| 12 | na | | | | .031667 | | | | | .047500 | | | .047500 | .095000 | .100000 |
| 12 | wod | | | | .009500 | | | | | .014250 | | | .014250 | .028500 | .030000 |
| 12 | wp | | | | .038000 | | | | | .057000 | | | .057000 | .114000 | .120000 |

**Appendix References**

Brennan, R. L. (1981). *Elements of Generalizability Theory.* Iowa City, IA: ACT Publications.

Cronbach, L. J. Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement, 57,* 373-399.

Hambleton, R. K., Jeager, R. M., Koretz, D., Linn, R. L., Millman, J., & Phillips, S. E. (1995). *Review of the Measurement Quality of the Kentucky Instructional Results Information System, 1991-1994.* Frankfort, KY: Office of Educational accountability, Kentucky General Assembly.

Hoffman, R. G., Koger, L., & Awbrey, A. (1997). *Changes in spelling, capitalization, punctuation, and subject/verb agreement skills under the Kentucky Education Reform Act* (LRS97-1). Frankfort, KY: Bureau of Learning Results Services, Kentucky Department of Education.

Medsker, G. J. (1998). *Determining the Relationship between Student Transience and KIRIS* School *Results: Are Schools with Transient Students Unfairly Impacted?* (HumRRO Report FR-WATSD-98-12). Radcliff, KY: Human Resources Research Organization.

Miller, M. David. (April, 1999). *Generalizability of Performance-Based Assessments at the School Level.* Paper presented at the annual meeting of the American Educational Research Association, Montreal, April, 1999.

Thacker, A. A., Koger, L. E., Hoffman, R. G., & Koger, M. E. (2000). *The Transition from KIRIS to CATS, Year 2: Instruction, Communication, and Perceptions at 31 Kentucky Schools* (FR-WATSD-99-23). Alexandria, VA: Human Resources Research Organization.

Yen, W. M. (1997). The technical quality of performance assessments: Standard Errors of Percents of Pupils Reaching Standards. *Educational Measurement: Issues and Practice, 16,* 5-15.